

Consultative Committee for Space Data Systems

**RECOMMENDATION CONCERNING SPACE
DATA SYSTEMS STANDARDS**

Reference Model for an Open Archival Information System (OAIS)

**CCSDS 650.0-W-3.0
WHITE BOOK**

April 15, 1998



DRAFT RECOMMENDATION: REFERENCE MODEL FOR AN OAIS

Dear Reader:

The following version of the OAIS Reference Model is designated CCSDS 650.0-W-3.0. This version, called White Book 3.0, reflects most of the updates, both specific and general, agreed to at the ISO Frascati International Workshop held in November, 1997. Comments are highly desirable.

Comments may be sent to:

Lou Reich
louis.i.reich@gsfc.nasa.gov

or

Don Sawyer
donald.sawyer@gsfc.nasa.gov

A new version, before formal public review, is planned for issue following the May 1998 meeting in Houston.

AUTHORITY

Issue:	White Book, Issue 3.0
Date:	October 15, 1997
Location:	Silver Spring, USA

This document, when it has been approved for publication by the Management Council of the Consultative Committee for Space Data Systems (CCSDS), will reflect the consensus of technical panel experts from CCSDS Member Agencies.

This document is published and maintained by:

CCSDS Secretariat
Program Integration Division (Code MG)
National Aeronautics and Space Administration
Washington, DC 20546, USA

FOREWORD

This document is a technical Recommendation for use in developing a consensus on what is required for an archive to provide permanent, or indefinite long-term, preservation of digital information. It may be useful as a starting point for a similar document addressing the indefinite long-term preservation of non-digital information.

This Recommendation establishes a common framework of terms and concepts which comprise an Open Archival Information System (OAIS). It allows existing and future archives to be more meaningfully compared and contrasted. It provides a basis for further standardization within an archival context and it should promote greater vendor awareness of, and support of, archival requirements.

Through the process of normal evolution, it is expected that expansion, deletion, or modification to this document may occur. This Recommendation is therefore subject to CCSDS document management and change control procedures, which are defined in Reference [1].

DOCUMENT CONTROL

Document	Title	Date	Status and Substantive Changes
CCSDS 650.0-W-1	Report Concerning Space Data Systems Standards: Reference Model for an Open Archival Information System (OAIS)	April 1997	Original Issue
CCSDS 650.0-W-1.1		July 1997	Revised information model and reduced text and complexity in Section 2. Partial response to Silver Spring Workshop directions.
CCSDS 650.0-W-1.2		Sept. 1997	Further revision of information model to incorporate concept of packaging information,. More complete response to Silver Spring Workshop directions.
CCSDS 650.0-W-2.0		Oct. 1997	Complete revisions re: Silver Spring Workshop instructions.
CCSDS 650.0-W-3.0		April 1998	Complete revisions re: Frascati workshop instructions

CONTENTS

<u>Section</u>	<u>Page</u>
1 INTRODUCTION	1
1.1 PURPOSE AND SCOPE	1
1.2 APPLICABILITY	2
1.3 RATIONALE	2
1.4 ROAD-MAP FOR DEVELOPMENT OF RELATED STANDARDS	3
1.5 DOCUMENT STRUCTURE	3
1.6 DEFINITIONS	5
1.6.1 ACRONYMS AND ABBREVIATIONS	5
1.6.2 TERMS	5
1.7 REFERENCES	12
2 OAIS CONCEPTS	13
2.1 OAIS ENVIRONMENT	14
2.2 OAIS INFORMATION	15
2.2.1 INFORMATION DEFINITION	15
2.2.2 INFORMATION PACKAGE DEFINITION	16
2.2.3 INFORMATION PACKAGE VARIANTS	17
2.3 OAIS HIGH LEVEL EXTERNAL INTERACTIONS	18
2.3.1 MANAGEMENT INTERACTION	19
2.3.2 PRODUCER INTERACTION	20
2.3.3 CONSUMER INTERACTION	20
2.3.4. VARIATIONS IN ARCHIVAL ENVIRONMENT	21
3 OAIS RESPONSIBILITIES	23
3.1 NEGOTIATES AND ACCEPTS INFORMATION	23
3.2 ASSUMES SUFFICIENT CONTROL FOR PRESERVATION	24
3.3 DETERMINES DESIGNATED CONSUMER COMMUNITIES	26
3.4 ENSURES INFORMATION IS INDEPENDENTLY UNDERSTANDABLE	26

3.5	FOLLOWS ESTABLISHED PRESERVATION POLICIES AND PROCEDURES	27
3.6	MAKES THE INFORMATION AVAILABLE	28
4	DETAILED MODELS	29
4.1	FUNCTIONAL MODEL	29
4.1.1	COMMON SERVICES	30
4.1.2	INGEST	30
4.1.3	ARCHIVAL STORAGE	31
4.1.4	DATA MANAGEMENT	33
4.1.5	ADMINISTRATION	33
4.1.6	ACCESS AND DISSEMINATION	34
4.2	INFORMATION MODEL	39
4.2.1	LOGICAL MODEL FOR ARCHIVAL INFORMATION	39
4.2.2	LOGICAL MODEL OF INFORMATION IN AN OPEN ARCHIVAL INFORMATION SYSTEM (OAIS)	50
4.2.3	DATA MANAGEMENT DATA	62
4.3	HIGH LEVEL DATA FLOWS AND TRANSFORMATIONS	64
4.3.1	DATA TRANSFORMATIONS IN THE PRODUCER ENTITY	65
4.3.2.	DATA TRANSFORMATIONS IN THE INGEST FUNCTIONAL AREA	65
4.3.3.	DATA TRANSFORMATIONS BY THE ARCHIVAL STORAGE AND DATA MANAGEMENT FUNCTIONAL AREAS	67
4.3.4.	DATA FLOWS AND TRANSFORMATIONS IN THE ACCESS/DISSEMINATION FUNCTIONAL AREA	67
5	DIGITAL MIGRATION PERSPECTIVES	69
5.1	DIGITAL MIGRATION MOTIVATORS	69
5.2	AIP OBJECTS : COMPONENTS, POINTERS, AND PACKAGE STRUCTURES	70
5.2.1	CONTENT INFORMATION AND PDI COMPONENTS AND POINTERS	70
5.2.2	PACKAGING INFORMATION STRUCTURES AND POINTERS	73
5.2.3	MODELING POINTERS	75
5.3	DIGITAL MIGRATION ENVIRONMENT MODEL	76
5.4	CATEGORIES OF PRESERVATION	77
5.4.1	MEDIA FORMAT PRESERVATION	78
5.4.2	MEDIA SUPPORTING-ATTRIBUTE PRESERVATION	78
5.4.3	MEDIA FILE CONTENT PRESERVATION	78
5.5	MIGRATION TYPES	79
5.5.1	REPACKAGING	79
5.5.2	REPACKAGING EXAMPLES	79

5.5.3	TRANSMUTATION	81
5.5.4	TRANSMUTATION EXAMPLES	81
5.6	DISTINGUISHING AIP VERSIONS, EDITIONS AND DERIVED AIPS	83
6	ARCHIVE ASSOCIATIONS	84
6.1.	TECHNICAL LEVELS OF INTERACTION BETWEEN OAIS ARCHIVES	84
6.1.1.	INDEPENDENT ARCHIVES	85
6.1.2.	COOPERATING ARCHIVES	85
6.1.3.	FEDERATED ARCHIVES	86
6.1.4.	ARCHIVES WITH SHARED FUNCTIONAL AREAS	88
6.2.	MANAGEMENT ISSUES WITH FEDERATED ARCHIVES	89
ANNEX A.	SCENARIOS OF EXISTING ARCHIVES	91
A.1	PLANETARY DATA SYSTEM ARCHIVE	91
A.2	NATIONAL ARCHIVES AND RECORDS ADMINISTRATION'S CENTER FOR ELECTRONIC RECORDS	95
A.3	LIFE SCIENCES DATA ARCHIVE	102
A.4	NATIONAL COLLABORATIVE PERINATEL PROJECT (NCPP) 1959-1974	107
A.5	ARCHIVE SCENARIO FOR THE <i>CENTRE DES DONNEES DE LA PHYSIQUE DES PLASMAS</i> (CDPP)	112
ANNEX B.	COMPATIBILITY WITH OTHER STANDARDS	119
ANNEX C.	BRIEF GUIDE TO THE OMT	120
ANNEX D.	ENTITY AND FUNCTION MATRIX	122
ANNEX D.	ILLUSTRATIVE SCENARIOS	123

1 INTRODUCTION

1.1 PURPOSE AND SCOPE

The purpose of this document is to define the ISO Reference Model for an **Open Archival Information System** (OAIS). An OAIS is an archive, consisting of an organization of people and systems, that has accepted the responsibility to preserve information and make it available for one or more designated communities. It meets a set of such responsibilities as defined in this document and this allows an OAIS archive to be distinguished from other uses of the term 'archive.' The term 'Open' in OAIS is used to imply that this standard and future related standards are developed in open forums, and it does not imply that access to the archive is unrestricted.

The information being maintained has been deemed to need **Long-Term Preservation**, even if the OAIS itself is not permanent. **Long Term** is long enough to be concerned with the impacts of changing technologies, including support for new media and data formats, and it extends indefinitely. In this reference model there is a particular focus on digital information, both as the primary forms of information held and as supporting information for both digitally and physically archived materials. Therefore the model accommodates information that is inherently non-digital (e.g. a physical sample) but the modeling and preservation of such information is not addressed in detail. This reference model:

- provides a framework for the understanding and increased awareness of archival concepts needed for long-term digital information preservation and access.
- provides the concepts needed by non-archival organizations to be effective participants in the preservation process.
- provides a framework that facilitates the description and comparison of the architectures and operations of existing and future archives.
- provides a basis for comparing the data models of digital information preserved by archives and for discussing how data models and the underlying information may change over time.
- provides a foundation that may be expanded by other efforts to cover long-term preservation of information that is NOT in digital form (e.g., physical media, physical samples).
- expands consensus on the elements and processes for long-term digital information preservation and access, and it promotes a larger market which vendors can support.
- guides the identification and production of OAIS related standards.

The Reference Model addresses a full range of archival information preservation functions

including ingest, archival storage, data management, access, and dissemination. It also addresses the migration of digital information to new media and forms, the data models used to represent the information, the role of software in information preservation, and the exchange of digital information among archives. It identifies both internal and external interfaces to the archive functions, and it identifies a number of high level services at these interfaces.

1.2 APPLICABILITY

The OAIS model in this document may be applicable to any archive. It is specifically applicable to organizations with the responsibility of making information available for the long term. This model is also of interest to those organizations and individuals who create information that may need long-term preservation and those that may need to acquire information from such archives.

This model may also be useful for those organizations developing or operating temporary archives for three reasons:

- When taking into consideration the rapid pace of technology, there is the likelihood that many archives, thought of as temporary, will in fact find that some or much of their holdings will need the same type of attention as that given by permanent archives.
- Although some archives may be temporary, some or all of their information may need to be preserved indefinitely. Such archives need to be active participants in the long-term preservation effort.
- Many of the functional and informational modeling concepts may be relevant to the comparison and design of temporary archives.

Standards developers are expected to use this model as a basis for further standardization and therefore provide an extension of what is meant by "operating an OAIS archive". A large number of related standards are possible. A road-map for such development is briefly addressed in Section 1.4.

This Reference Model does not specify an implementation. Actual implementations may group or break out functionality differently.

1.3 RATIONALE

A tremendous growth in computational power, and in networking bandwidth and connectivity, has resulted in an explosion of organizations who are making information available in electronic forms. Transactions among all types of organizations are being conducted using electronic forms that are taking the place of more traditional forms such as paper.

Preserving information in electronic forms is much more difficult than for forms such as

paper and film. This is not only a problem for traditional archives, but for many organizations that have never thought of themselves as performing an archival function. It is expected that this reference model, by establishing minimum requirements for an OAIS archive along with a set of archival concepts, will provide a common framework from which to view archival challenges, particularly as they relate to digital information. This should enable more organizations to understand the issues and to take proper steps to ensure long term information preservation. It should also provide a basis for more standardization and therefore a larger market that vendors can support in meeting archival requirements.

1.4 ROAD-MAP FOR DEVELOPMENT OF RELATED STANDARDS

This Reference Model serves to identify areas suitable for the development of OAIS related standards. Some of these standards may be developed by Panel 2 of the CCSDS (sub-committee of ISO); others may be developed by other standardization bodies. However, any such work undertaken by other bodies should be coordinated with CCSDS Panel 2 in order to minimize incompatibilities and efforts. Areas for potential OAIS related standards include:

- standard(s) for the interfaces between OAIS type archives.
- standard(s) for the submission (ingest) of digital data sources to the archive.
- standard(s) for the delivery of digital sources from the archive.
- standard(s) for the submission of digital metadata, about digital or physical data sources, to the archive. Here, it can be envisaged that different disciplines might require different metadata standards.
- standard(s) for the identification of digital data within the archive
- protocol standard(s) to search and retrieve metadata information about digital and physical data sources.
- standard(s) for media access allowing replacement of media management systems without having to re-write the media
- standard(s) for specific physical media
- standard(s) for the migration of information across media and representations
- standard(s) for recommended archival practices and accreditation of archives

1.5 DOCUMENT STRUCTURE

Section 2 provides a high level overview of the major concepts involved in an OAIS archive. It provides a view of the environment of an OAIS archive and the roles played by those who interact with it. It discusses what is meant by “information” and what is necessary to preserve it for the long term.

Section 3 defines mandatory responsibilities an OAIS archive must discharge in preserving its information.

Section 4 provides model views needed for a detailed understanding of an OAIS archive. It breaks down the OAIS into a number of functional areas and it identifies some high level

services at the interfaces. It also provides detailed data model views of information using OMT diagrams.

Section 5 provides some perspectives on the issue of migration of information across media and across new formats or representations.

Section 6 is an introduction to the various alternatives for archives to associate with other archives to provide increased services.

Annex A provides scenarios of existing archive operations.

Annex B relates parts of this reference model to other standards work

Annex C. provides a brief tutorial on the Object Modeling Technique (OMT).

Annex D maps the relationships of functions and entities within the data flow model of the OAIS..

Annex E provides some preliminary work on scenarios, using the terms and concepts defined previously, to show how information may flow into and out of an archive, and how information may be migrated within an archive.

1.6 DEFINITIONS

1.6.1 ACRONYMS AND ABBREVIATIONS

AIC - Archival Information Collection
AIP - Archival Information Package
AIU - Archival Information Unit
ASCII - American Standard Code for Information Interchange
CAD - Computer-Automated Design
CCSDS - Consultative Committee for Space Data Systems
CD-ROM - Compact Disk - Read Only Memory
CIP - Catalog Inter-operability Protocol
CRC - Cyclical Redundancy Check
CT - Computer Tomography
DED - Data Entity Dictionary
DBMS - Data Base Management System
DDL - Data Description Language
DIP - Dissemination Information Package
DVD - Digital Video Disk
EBCDIC - Extended Binary Coded Decimal Interchange Code
FITS - Flexible Image Transfer System
GIF - Graphics Interchange Format
HFMS - Hierarchical File Management System
IEEE - Institute of Electrical and Electronic Engineers
ISBN - International Standard Book Number
ISO - Organization for International Standardization
LSDA - Life Sciences Data Archive
NARA - National Archives and Records Administration
NASA - National Aeronautics and Space Administration
NSSDC - National Space Science Data Center
OAIS - Open Archival Information System
ODL - Object Description Language
OMT - Object Modeling Technique
PDI - Preservation Description Information
PDMP - Project Data Management Plan
PDS - Planetary Data System
PSDD - Planetary Science Data Dictionary
SIP - Submission Information Package
UNICODE - Universal Code
WWW - World-Wide Web

1.6.2 TERMS

Access: This OAIS entity contains the services and functions which make the archival information and externally-available services visible to Consumers.

Access Aids: Software or documents that allow Consumers to locate, analyze, and order Archival Information Packages of interest.

Access Methods: A method for retrieving an Archival Information Package based on its name or identifier which is available to authorized users.

Adhoc Order: A request that is generated by a Consumer for information the OAIS has indicated is currently available.

Administration: This OAIS entity contains the services and functions needed to control the operation of the other OAIS functional entities on a day to day basis

Archive: A organization that intends to preserve information for access and use by one or more Designated Communities.

Archival Storage: This OAIS entity contains the services and functions used for the storage and retrieval of Archival Information Packages.

Archival Information Collection (AIC): An Archival Information Package whose Content Information is an aggregation of other Archival Information Packages.

Archival Information Package (AIP): An information packaging concept that requires the presence of Content Information and all the associated Preserving Description Information that is needed to preserve the Content Information over the long term. It has associated Packaging Information.

Archival Information Unit (AIU): An Archival Information Package whose Content Information is not further broken down into other Content Information components, each of which has its own complete Preservation Description Information. It can be viewed as an “atomic” AIP. An example of an AIU would be a table of numbers representing temperatures in a certain region with all the associated documentation describing how and where the temperatures were measured, what instruments were used to make the measurements, who made the measurements, why they were made, what processing has been performed on the measurements and who has had custody of these measurements since they were first created, how the measurements relates to other information, how the measurement can be uniquely referenced by others, etc.

Associated Descriptions: Information describing the content of an Information Package from the point of view of a particular Access Aid.

Client: An application which exchanges information with another application (see also Consumer).

Collection Description: An Associated Description that described the collection as a whole.

Collection Descriptor: A type of Package Descriptor that is specialized to provide

searchable information about a collection.

Common Services: Supporting services such as inter-process communication, name services, temporary storage allocation, exception handling, security, and directory services necessary to support the OAIS.

Component: An independent, separately identifiable, part of a Content Information or PDI Information Object

Consumer: The role played by those persons, or client systems, who interact with OAIS services to find preserved information of interest and to access that information in detail.

Content Information (CI): That set of information that is the primary target for preservation. It is distinguished from Preservation Description Information which is used to assist in the preservation of the Content Information. An example of Content Information could be a single table of numbers representing, and understandable as, temperatures but it excludes the documentation which would explain its history and origin, how it relates to other observations, etc.

Context Information: Information that documents the relationships of the Content Information to its environment. This includes why the Content Information was created, and how it relates to other Content Information objects existing elsewhere.

Data: The representation forms of information. Examples of data include a sequence of bits, a table of numbers, the characters on a page, the sounds made by a person speaking, a moon rock.

Data Delivery Session: May be a delivered set of media or a single telecommunications session. The Data Delivery Session format/contents is based on a data model negotiated between the OAIS and the producer in the Submission Agreement. This data model identifies the logical constructs used by the producer and how they are represented on each media delivery or in the telecommunication session.

Data Dissemination Session: May be a delivered set of media or a single telecommunications session. The Data Dissemination Session format/contents is based on a data model negotiated between the OAIS and the Consumer in the Request Agreement. This data model identifies the logical constructs used by the OAIS and how they are represented on each media delivery or in the telecommunication session.

Data Dictionary: A formal repository of terms used to describe data.

Data Management: This OAIS entity contains the services and functions for populating, maintaining, and querying a wide variety of information such as catalogs and inventories on what may be retrieved from Archival Storage, processing algorithms that may be run on retrieved data (if any), consumer access statistics, security controls, OAIS schedules and procedures.

Data Management Data: Data created and stored in Data Management persistent storage that refer to operation of an archive. Some examples of this data are accounting data for consumer billing and authorization, policy data, subscription data for repeating requests, and statistical data for generating reports to archive management.

Data Object: Either a Physical Object or a Digital Object.

Designated Community: An identification of a set of potential Consumers who should be able to understand a particular set of information.

Descriptive Information: That set of information, consisting primarily of Package Descriptors, which is provided to Data Management to support the finding of preserved information by Consumers.

Digital Object: An object composed of a set of bit sequences.

Dissemination: The act of providing the Dissemination Information Package to the Consumer.

Dissemination Information Package (DIP): An Information Package that contains parts or all of one or more AIPs, that is distributed to the Consumer as requested.

Embedded Pointer: A Pointer that is found within a Component.

Finding Aid: A type of Access Aid that allows a user to search for and identify Archival Information Packages of interest.

Fixity Information: This information documents the authentication mechanisms and provides authentication keys to ensure that the Content Information object has not been altered in an undocumented manner. An example is the CRC code for a file.

Format: The sequential organization of data in terms of its components.

Independently Usable Information: Information with sufficient documentation to allow the data to be understood and used by the Designated Community without having to resort to special resources not widely available, including named individuals.

Information: Any type of knowledge that can be exchanged. In an exchange, it is represented by data. Often the representation used is not fully known to the recipient of the data and the data must be accompanied by explicit Representation Information, understandable to the recipient, that is used to interpret the data. An example is a string of bits (the data) accompanied by a description of how to interpret a string of bits as numbers representing temperature observations measured in degrees Celsius (the representation information).

Information Object (IO): A Data Object together with optional Representation Information.

Information Package: An information packaging concept that distinguishes Content Information from associated Preservation Description Information where the Preservation Description Information applies to the Content Information and is needed to aid in the preservation of the Content Information. It has associated Packaging Information used to delimit and identify the Content Information and Preservation Description Information.

Ingest: This entity contains the services and functions that accept Submission Information Packages from Producers, prepare Archival Information Packages for storage, and ensure that Archival Information Packages and their supporting Descriptive Information become established within the OAIS.

Long Term: A period of time which is long enough to be concerned about the impacts of changing technologies, including support for new media and data formats, on the information being held in a repository. This period extends into the indefinite future.

Long-term Preservation: The act of preserving information, in a form which can be made understandable to a Designated Community, over the Long Term.

Management: Management is the role played by those who set overall OAIS policy as one component in a broader policy domain.

Member Description: An Associated Description that describes a member of a collection.

Metadata: Data about other data.

Migration: The act of transferring a digital Information Object to new storage media or to new forms.

Open Archival Information System (OAIS): An OAIS is an archive, consisting of an organization of people and systems, that has accepted the responsibility to preserve information and make it available for one or more designated communities. It meets a set of such responsibilities as defined in Section 3 of this document and this allows an OAIS archive to be distinguished from other uses of the term 'archive.' The term 'Open' in OAIS is used to imply that this standard and future related standards are developed in open forums, and it does not imply that access to the archive is unrestricted.

Order Agreement: An agreement between the archive and the Consumer in which the physical details of the delivery such as media type and object format are specified.

Package Structure: Information that encapsulates and identifies Components and Representation Pointers.

Packaging Information : That information that is used to bind and identify the components

of an Information Package. For example, it may be the ISO-9660 volume and directory information used on a CD-ROM to provide the content of several files containing Content Information and Preservation Description Information. It is comprised of Package Structures and Package Pointers.

Physical Object: An object (such as a moon rock, biospecimen, microscope slide) with physically observable properties that represent information that is considered suitable for being adequately documented for preservation, distribution and independent usage.

Pointer: A name, which may be segmented, that identifies one or more Components or Package Structures.

Preserve: Maintain information, in a correct and Independently Usable form, over the Long Term.

Preservation Description Information (PDI): Information necessary to adequately preserve the Content Information and which can be categorized as Provenance, Reference, Fixity, and Context information.

Producer. The role played by those persons, or client systems, who provide the information to be preserved.

Provenance Information: Information that documents the history of the Content Information. This information tells the origin or source of the Content Information, any changes that may have taken place since it was originated, and who has had custody of it since it was originated. Examples of Provenance Information are the principal investigator who recorded the data and the information concerning its storage, handling and migration.

Reference Information: Information that identifies, and if necessary describes, one or more mechanisms used to provide assigned identifiers for the Content Information. It also provides these identifiers that allow outside systems to refer, unambiguously, to this particular Content Information. An example of reference information is an ISBN.

Reference Model: A framework for understanding significant relationships among the entities of some environment, and for the development of consistent standards or specifications supporting that environment. A reference model is based on a small number of unifying concepts and may be used as a basis for education and explaining standards to a non-specialist.

Repackaging: A type of migration in which there is an alteration in the Packaging Information of the AIP.

Replication: When moving a digital Information Object to new media, do a bit-for-bit copy so that the new media is identical to the old media.

Representation Information: Information that maps data into more meaningful concepts.

An example is the ASCII definition which describes how bits (i.e., data) are mapped into numbers. Another example is a description of the numbers (i.e., data) of a table as being the coordinates of a location on the Earth measured in East longitude and latitude.

Representation Net: The set of Representation Information which fully describes the meaning of a Data Object. Representation Information in digital forms needs additional Representation Information so its digital forms can be understood over the Long Term.

Representation Pointer: A Pointer which points to Representation Information from a Package Structure environment.

Result Set: The set of descriptive records for those AIPs in an OAIS which match the criteria stated in a Consumer query., or other results from a search on Data Management.

Search Session: A session initiated by the Consumer with the archive during which the Consumer will use the archive Finding Aids to identify and investigate potential holdings of interest.

Semantic Information: Information which imparts meaning apart from how other information is organized.

Structure Information: Information that imparts meaning about how other information is organized. For example, it maps bit streams to common computer types such as characters, numbers, and pixels and aggregations of those types such as character strings and arrays.

Submission Agreement: An agreement reached between an archive and the Producer that negotiates a data model for the Data Delivery Session. This data model identifies format/contents and the logical constructs used by the Producer and how they are represented on each media delivery or in the telecommunication session. It also transfers legal and physical custody of the data to the archive and the terms defining restrictions and access.

Submission Information Package (SIP): The Information Package identified by the Producer in the Submission Agreement with the OAIS

Subscription Order: A request that is generated by a Consumer for information that is to be delivered periodically on the basis of some event or events.

Transmutation: A type of Migration in which there is an alteration to the Content Information of an Archival Information Package. For example, changing ASCII codes to UNICODE in a text document being preserved..

Unit Descriptor: A type of Package Descriptor that is specialized to provide searchable information about an Archival Information Unit.

Universal Identifier: An identifier that always has an associated mapping to information that may be used to locate a Component. The mapping

1.7 REFERENCES

- [1] "Preserving Digital Information", Report of the Task Force on Archiving of Digital Information, final report currently available from the URL <<http://www.rlg.org/ArchTF/>>, May 1, 1996
- [2] "Object-Oriented Modeling and Design", by Rumbaugh, J., Blaha, M., Premerlani, W., Eddy, F., and Lorensen, W., Prentice Hall, 1991.
- [3] "Z39.50 Profile for Access to Digital Collections", currently available at the URL <<http://vinca.cnidr.org/protocols/profiles/zdl.html>>

2 OAIS CONCEPTS

The purpose of this section is to motivate and describe several key high level OAIS concepts. A more detailed view, and a formal modeling of these concepts, is given in Section 4.

The term “archive” has come to be used to refer to a wide variety of storage and preservation functions and systems. Traditionally, an archive is understood as a facility or organization which preserves records, originally generated by or for a government organization, institution, or corporation, for access by public or private communities. It accomplishes this task by taking ownership of the records, ensuring that they are understandable to the accessing community, and managing them so as to preserve their information content and authenticity. Historically, these records have been in such forms as books, papers, maps, photographs, and film which can be read directly by humans, or read with the aid of simple optical magnification and scanning aids. The major focus for preserving this information has been to ensure that they are on media with long term stability and that access to this media is carefully controlled.

The explosive growth of information in digital forms has posed a severe challenge not only for traditional archives and their information providers, but for many other organizations in the government, commercial and non-profit sectors. These organizations are finding, or will find, that they need to take on the information preservation functions typically associated with traditional archives because digital information is easily lost or corrupted. The pace of technology evolution is causing some hardware and software systems to become obsolete in a matter of a few years, and these changes can put severe pressure on the ability of the related data structures or formats to continue effective representation of the full information desired. Although some archives may be temporary, some or all of their information may need to be preserved indefinitely. Much of the supporting information necessary to preserve this information is more easily available or only available at the time when the original information is produced. Such temporary archives need to be active participants in the long-term preservation effort and follow the principles espoused in this OAIS reference model to ensure the information can be preserved for the long term. Participation in these efforts will minimize the lifecycle costs and enable effective long-term preservation of the information

A major purpose of this reference model is to facilitate a much wider understanding of what is required to preserve and access information for the Long Term. To avoid confusion with simple “bit storage” functions, the reference model defines an Open Archival Information System (OAIS) which performs a Long-term information preservation function. An OAIS archive is one that intends to preserve information for access and use by one or more designated communities, and it meets the requirements given in Section 3. It includes archives that have to keep up with steady input streams of information as well as those that experience primarily aperiodic inputs. It includes archives that provide a wide variety of sophisticated access services as well as those that support only the simplest types of requests. For the remainder of this document, the term archive is understood to refer to an OAIS, or OAIS archive, unless the context makes it clear otherwise (e.g., traditional archives).

The OAIS model recognizes the already highly distributed nature of digital information

holdings and the need for local implementations of effective policies and procedures supporting information preservation. This allows, in principle, a wide variety of organizational arrangements, including various roles for traditional archives, in achieving this preservation. It is expected that organizations attempting to preserve information will find that using OAIS terms and concepts will assist them in achieving their information preservation goals.

At the same time, the problems associated with achieving economical and truly effective Long-term digital information preservation should not be underestimated. A good survey of many of the issues is contained in a report by the Task Force on Archiving of Digital Information entitled “Preserving Digital Information” [Reference 1].

2.1 OAIS ENVIRONMENT

The environment surrounding an OAIS is given by the simple model shown in **Figure 2-1**

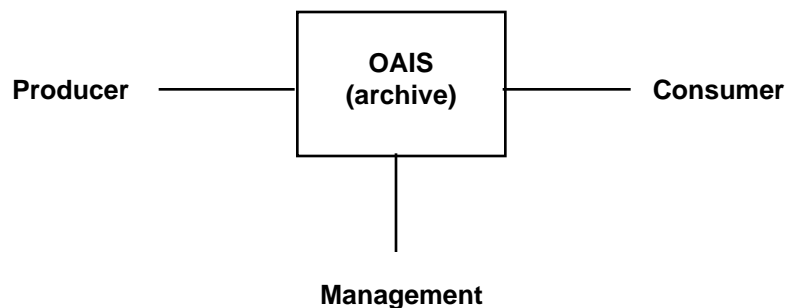


Figure 2-1. Environment Model of an OAIS

Outside the OAIS are **Producers**, **Consumers**, and **Management**. Producer is the role played by those persons, or client systems, who provide the information to be preserved. Management is the role played by those who set overall OAIS policy as one component in a broader policy domain. In other words, Management control of the OAIS is only one of Management’s responsibilities. Management is not involved in day-to-day archive operations. The responsibility of managing the OAIS on a day-to-day basis is included within the OAIS in an administrative function and will be described in Section 4. Consumer is the role played by those persons, or client systems, who interact with OAIS services to find and acquire preserved information of interest.

Other OAIS archives are not shown explicitly. Such archives may establish particular agreements among themselves consistent with Management and OAIS needs. Other archives may interact with a particular archive for a variety of reasons and with varying degrees of formalism for any pre-arranged agreements. One OAIS may take the role of Producer to another OAIS; an example is when the responsibility for preserving a type of information is to be moved to this other archive. One OAIS may take the role of Consumer to another OAIS; an example is when the first OAIS decides to rely on the other OAIS for a type of information it seldom needs and chooses not to preserve locally. Such reliance should have

some formal basis that includes the requirement for communication between the archives of any policy changes which might affect this reliance.

2.2 OAIS INFORMATION

2.2.1 INFORMATION DEFINITION

A clear definition of information is central to the ability of an OAIS to preserve it. While formal modeling of information is given in Section 4, some key concepts are given in this section.

A person, or system, can be said to have a knowledge base which allows them to understand received information. For example, a person who has a knowledge base that includes an understanding of English will be able to read, and understand, an English text.

Information is defined as any type of knowledge that can be exchanged, and this information is always expressed (i.e., represented) by some type of data. For example, the information in a hardcopy book is typically expressed by the observable characters (the data) which, when they are combined with a knowledge of the language used (the knowledge base), are converted to more meaningful information. If the recipient does not already include English in its knowledge base, then the English text (the data) need to be accompanied with an English dictionary and grammar information (i.e., Representation Information) in a form that is understandable using the recipients knowledge base.

Similarly, the information stored within a CD-ROM file is expressed by the bits (the data) it contains which, when they are combined with the Representation Information for those bits, are converted to more meaningful information as long as the Representation Information is understandable using the recipients knowledge base. For example, assume the bits represent an ASCII table of numbers giving the coordinates of a location on the Earth measured in degrees latitude and East longitude. The Representation Information will typically include the definition of ASCII together with descriptions of the format of the numbers and their locations in the file, their definitions as latitude and longitude, and the definition of their units as degrees. It may go on to include additional meaning that is assigned to the table.

In general, it can be said that “Data interpreted using its Representation Information yields Information” and this is shown schematically in **Figure 2-2**.

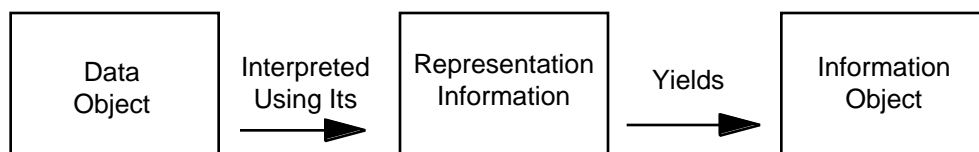


Figure 2-2. Obtaining Information from Data

In order for this **Information Object** to be successfully preserved, it is critical for an archive to clearly identify and understand the **Data Object** and its associated Representation

Information. For digital information, this means the archive must clearly identify the bits and the Representation Information that applies to those bits. This required transparency to the bit level is a distinguishing feature of digital information preservation and it runs counter to the information hiding trend which is so successful in supporting modern computing services. This presents a significant challenge to the preservation of digital information. (Note: the recursive nature of Representation Information, which typically is composed of its own data and other Representation Information, is dealt with in Section 4.)

This definition of an Information Object is applicable to all the information types discussed in the following sections. In other words, they all have associated Representation Information although this is usually not shown explicitly.

2.2.2 INFORMATION PACKAGE DEFINITION

Every submission of information to an OAIS by a Producer, and every dissemination of information to a Consumer, occurs as one or more discrete transmissions. Therefore it is convenient to define the concept of an Information Package (IP).

An **Information Package (IP)** is a conceptual container of two types of information called **Content Information** and **Preservation Description Information (PDI)**. The Content Information and PDI are viewed as being encapsulated and identifiable by the **Packaging Information**. The resulting package is viewed as being discoverable by virtue of the **Descriptive Information**.

These Information Package relationships are shown schematically in **Figure 2-3**.

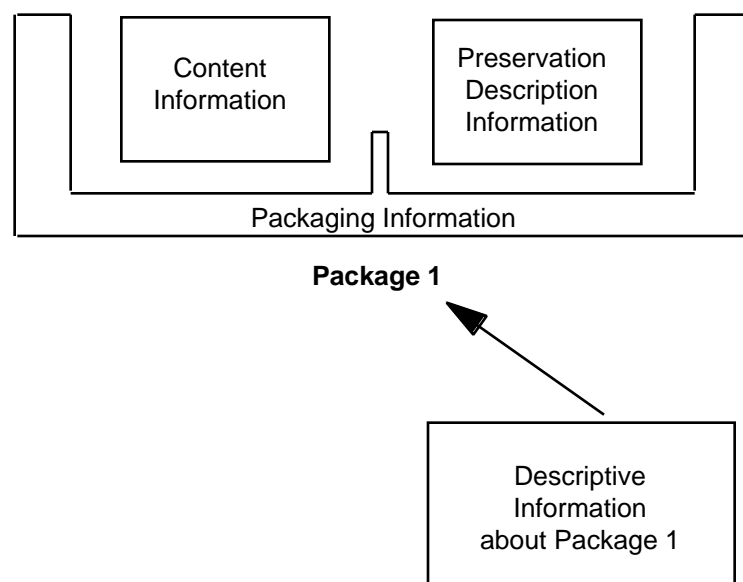


Figure 2-3. Information Package Concepts and Relationships

The Content Information is that information which is the primary target of preservation. It consists of the primary Digital Object (i.e., bits) and its associated Representation Information needed to make the Digital Object understandable. For example, it may be an image provided as the bit content of a CD-ROM file together with the Representation Information for those bits.

Only after the Content Information has been clearly defined can an assessment of the Preservation Description Information be made. The Preservation Description Information applies to the Content Information and is needed to make the Content Information preservable over the Long Term. The Preservation Description Information is divided into four types of preserving information called Provenance, Context, Reference, and Fixity. Briefly, they are the following:

- Provenance describes the source of the Content Information, who has had custody of it since its origination, and what its history (including processing history) has been.
- Context describes how the Content Information relates to other information outside the Information Package. For example, it would describe why the Content Information was produced and it may include a description of how it relates to another Content Information object that is available.
- Reference provides one or more identifiers, or systems of identifiers, by which the Content Information may be uniquely identified. Examples include an ISBN number for a book or a set of attributes that distinguish one Content Information from another.
- Fixity provides a wrapper, or protective shield, that protects the Content Information from unintended alteration. For example, it may involve a check sum over the Content Information of a digital Information Package.

The Packaging Information is that information which, either actually or logically, binds and relates the Content Information and PDI into an entity. For example, if the Content Information and PDI are identified as being the content of specific files on a CD-ROM, then the Packaging Information would include the ISO-9660 volume/file structure on the CD-ROM as well as the physical CD-ROM disk.

The Descriptive Information is that information which is used to discover which package has the Content Information of interest. Depending on the setting, this may be no more than a descriptive title of the Information Package that appears in some message, or it may be a full set of attributes that are searchable in a catalog service.

2.2.3 INFORMATION PACKAGE VARIANTS

It is necessary to distinguish between an Information Package that is preserved by an OAIS and the Information Packages that are submitted to, and disseminated from, an OAIS. These variant packages are needed to reflect the reality that some submissions to an OAIS will have insufficient Representation Information or PDI to meet final OAIS preservation requirements. In addition, they may be organized very differently from the way the OAIS

organizes the information it is preserving. Finally, the OAIS may provide information to Consumers that does not include all the Representation Information or all the PDI with the associated Content Information being disseminated. These variants are referred to as the Submission Information Package (SIP), the Archival Information Package (AIP), and the Dissemination Information Package (DIP). Although they are all information packages, they differ in what is mandatory content.

The **Submission Information Package (SIP)** is that package that is sent to an OAIS by a Producer. Its form and detailed content is typically negotiated between the Producer and the OAIS. Most SIPs will have some Content Information and some PDI, but it may require several SIPs to provide a complete set of Content Information and associated PDI. The Content Information and the PDI both have associated Representation Information, and if there are multiple SIPs involved that use the same Representation Information, it is likely that such Representation Information will only be provided once to the OAIS. As another variation, since some types of PDI will apply to multiple SIPs from the same source, such PDI may be provided in a separate SIP that is without Content Information. The Packaging Information will always be present in some form. The Descriptive Information associated with a SIP is likely to be provided prior to submitting the SIP to the OAIS, but it may be provided at any time. It may be no more than a text description with a name or title, carried by the Packaging Information, by which the SIP may be recognized.

Within the OAIS one or more SIPs is transformed into one or more **Archival Information Packages (AIP)** for preservation. The AIP has a complete set of PDI for the associated Content Information. The AIP may also contain a collection of other AIPs and this is discussed and modeled in Section 4. The Packaging Information of the AIP will conform to OAIS internal standards, and it may vary as it is managed by the OAIS. The Descriptive Information associated with an AIP may be extensive and will be managed by the OAIS so that Consumers can find and order the Content Information of interest.

In response to an Order, the OAIS provides all or a part of an AIP to a Consumer in the form of a **Dissemination Information Package (DIP)**. The DIP may also include collections of AIPs, and it may or may not have complete PDI. The Packaging Information will always be present in some form so that the Consumer can clearly distinguish the information Ordered. The Packaging Information may take several forms depending on the dissemination media and Consumer requirements. The Descriptive Information associated with a DIP may be provided with the transfer of the DIP, or it may be provided at any time before or after the transfer. Its purpose is to give the Consumer enough information to recognize the DIP from among possible similar packages. It may be no more than a text description with a name or title, as carried by the Packaging Information, by which the DIP may be recognized.

2.3 OAIS HIGH LEVEL EXTERNAL INTERACTIONS

The following sections present a high level view of the interaction between the entities identified in the OAIS environment. **Figure 2-5** is a data flow diagram that represents the operational OAIS archive external data flows. This diagram concentrates on the flow of information among Producers, Consumers and the OAIS and does not include flows that

involve Management. These flows are dealt with further in Section 4.

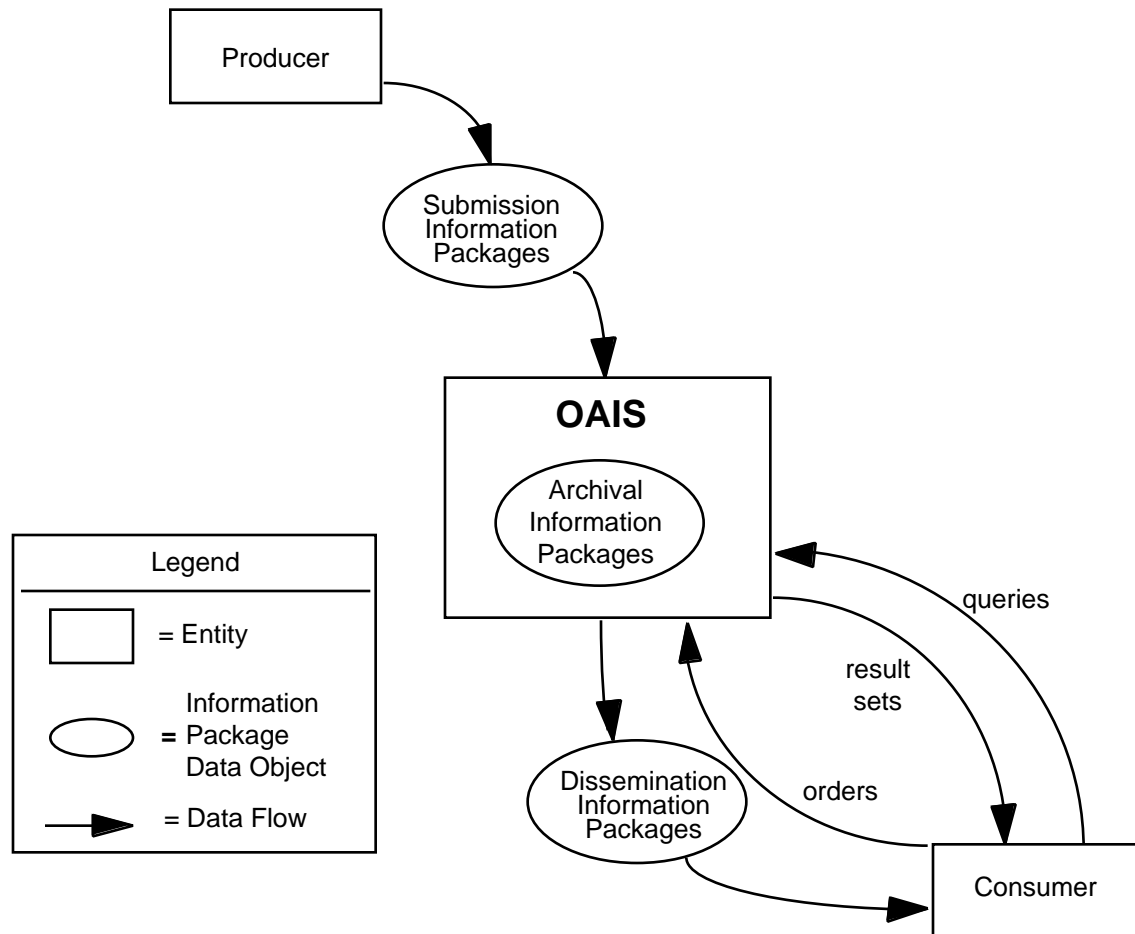


Figure 2-5. OAIS Archive External Data Flows

2.3.1 MANAGEMENT INTERACTION

- Management provides the OAIS with its charter and scope. The charter may be developed by the archive but it is important that Management formally endorse archive activities. The scope determines the breadth of both the Producer and Consumer groups served by the archive.
- Management is often the primary source of funding for an OAIS and may provide guidelines for resource utilization (personnel, equipment, facilities).
- Management will generally conduct some regular review process to evaluate OAIS performance and progress toward long-term goals.
- Management determines or at least endorses pricing policies, as applicable, for OAIS

services.

- Management participates in conflict resolution involving Producers, Consumers and OAIS internal administration.
- Management should also provide support for the OAIS by establishing procedures that assure OAIS utilization within its sphere of influence. For example, management policies should require that all funded activities within its sphere of influence submit data products to the archive and also adhere to archive standards and procedures.

2.3.2 PRODUCER INTERACTION

The Producer establishes a **Submission Agreement** with the OAIS, which identifies the SIPs to be submitted and may span any length of time for this submission. Some Submission Agreements will reflect a mandatory requirement to provide information to the OAIS, while others will reflect a voluntary offering of information. Within the Submission Agreement, one or more **Data Delivery Sessions**, usually with significant time gaps between the sessions, are recognized. A Data Delivery Session will contain one or more SIPs and may be a delivered set of media or a single telecommunications session. The Data Delivery Session content is based on a data model negotiated between the OAIS and the Producer in the Submission Agreement. This data model identifies the logical components of the SIP (e.g., the Content Information, PDI, Packaging Information, and Descriptive Information) that are to be provided and how (and whether) they are represented in each Data Delivery Session. All data deliveries within a Submission Agreement are recognized as belonging to that Submission Agreement and will generally have a consistent data model which is specified in the Submission Agreement. For example, a Data Delivery Session consists of a set of Content Information corresponding to a set of observations which are carried by a set of files on a CD-ROM. The Preservation Description Information is split among two other files. All of these files need Representation Information which must be provided in some way. The CD-ROM and its directory/file structure are the Packaging Information which provides encapsulation and identification of the Content Information and PDI. The Submission Agreement indicates how the Representation Information for each file is to be provided, how the CD-ROM is to be recognized, how the Packaging Information is used to encapsulate the Content Information and PDI, and it gives the frequency of Data Delivery Sessions (e.g. one per month for two years). It also gives other needed information such as access restrictions to the data.

Each SIP in a Data Delivery Session is expected to meet minimum OAIS requirements for completeness. However multiple SIPs may need to be received before an acceptable AIP is formed and fully ingested within the OAIS. At this point the information contained therein can become available, in principle, to OAIS Consumers. A Submission Agreement also includes, or references, the procedures and protocols by which an OAIS will either verify the arrival and completeness of a Data Delivery Session with the Producer or question the Producer on the contents of the Data Delivery Session.

2.3.3 CONSUMER INTERACTION

There are two common order types initiated by Consumers, the **Subscription Order** and the **Adhoc Order**.

- **Subscription Order:** The Consumer establishes an **Order Agreement** with the OAIS for information expected to be received periodically on the basis of some triggering event. It may span any length of time and under it one or more **Data Dissemination Sessions**, usually with significant time gaps between the sessions, may take place. A Data Dissemination session may involve the transfer of a set of media or a single telecommunications session. The Order Agreement identifies the components of one or more AIPs to be provided, identifies how they are mapped into a Dissemination Information Package (DIP) and how that DIP will be packaged in each media delivery or telecommunications session. The Order Agreement will also specify other needed information such as the trigger (e.g. event or time period) for new Data Dissemination Sessions, the criteria for selecting the OAIS holdings to be included in each new Data Dissemination Session, delivery information (e.g., name or mailing address), and any pricing agreements as applicable.
- **Adhoc Order:** The Consumer establishes an Order Agreement with the OAIS for information available from the archive. If the Consumer does not know a priori what specific holdings of the OAIS are of interest, the Consumer will establish a **Search Session** with the OAIS. During this Search Session the Consumer will use the OAIS **Finding Aids**, which operate on **Descriptive Information**, or in some cases on the AIPs themselves, to identify and investigate potential holdings of interest. This may be accomplished by the submission of queries and the return of result sets to the Consumer. This searching process tends to be iterative with a Consumer first identifying broad criteria and then refining this criteria based on previous search results. Once the Consumer identifies the OAIS AIPs, or AIP components, to be acquired, he may provide an Order Agreement to the OAIS to document the details of what components he will acquire and how he will acquire them. If the AIPs are available, an Adhoc Order will be placed. However if the AIPs desired are not yet available, a Subscription Order may be placed.

The Adhoc Order and the Subscription Order are similar although the Order Agreement may be substantially simpler with most Adhoc Orders. Note that the concept of an Order Agreement does not specify any particular implementation. It may, in some cases, be no more than the completion of a World Wide Web form.

2.3.4. VARIATIONS IN ARCHIVAL ENVIRONMENT

The explosion of computer processing power and digital media has resulted in many systems where the lines between producer, consumer and archive have been blurred. A good example of this is an **Active Archive**. In an Active Archive data is flowing regularly into the archive over an extended period of time at a rate driven by the Producer. This data is rapidly made accessible to Consumers. The Active Archive may also take the role of producer and be responsible for creation of products to be preserved based on the input data. Though the entire contents of this OAIS reference model is applicable to Active Archives, there are several additional factors facing an Active Archive including:

- The incoming data rate is not controlled by the archive. In active archives, this rate can be a major design factor.
- Due to the requirement of rapid access, preliminary versions of AIPs may be made available to the consumer community. These preliminary version may be updated before the final AIP is created so version management becomes a major policy/design issue.
- There will be an inherent conflict for available computer resources between the creation of AIPs and creation of specialized DIPs for the consumer

3 OAIS RESPONSIBILITIES

This section establishes mandatory responsibilities that an organization must discharge in order to operate an OAIS archive. The OAIS must:

- Negotiate and accept information from information producers.
- Assume sufficient control of the information provided to the level needed to ensure Long-term preservation.
- Determine (dependently or independently) which communities need to be able to understand the information provided.
- Ensure the information to be preserved is independently understandable to the **Designated Communities**. In other words, the communities should be able to understand the information without needing the assistance of the experts who produced the information.
- Follow documented policies and procedures which ensures the information is preserved against all reasonable contingencies and enables the information to be disseminated as authenticated copies of the original or as traceable to the original.
- Make the preserved information available to the designated communities..

In the following sections, each of these responsibilities is explored in greater depth for clarification.

3.1 NEGOTIATES AND ACCEPTS INFORMATION

An organization operating an OAIS will have established some criteria that aids in determining the types of information that it is willing to, or it is required to, accept. These criteria may include, among others, subject matter, information source, degree of uniqueness or originality, and the nature of the techniques used to represent the information (e.g., physical media, digital media, format). The information may, in general, be submitted using a wide variety of common and not-so-common forms, such as books, documents, maps, data sets, and moon rocks using a variety of communication paths including networks, mail, and special delivery.

The OAIS negotiates with the Producer to ensure it acquires the right Content Information and associated PDI for its mission and Consumer communities. It needs to extract, or otherwise obtain, sufficient Descriptive Information to assist its Consumer communities in finding the Content Information of interest. It also needs to ensure that the information meets all OAIS internal standards.

3.2 ASSUMES SUFFICIENT CONTROL FOR PRESERVATION

In general, the OAIS will accept the Content Information as either a custodian or as the new legal owner. When acting as a custodian, the OAIS may need to involve the actual owner(s) in some migration and access decisions depending on the authority it has been granted to act independently. When it is the legal owner, it already has the independence to do what is needed to preserve the information and make it accessible.

The OAIS must assume sufficient control over the Content Information and Preservation Description Information so that it is able to preserve it for the Long Term. There is no issue with the AIP's Packaging Information because, by definition, this is under internal OAIS control. The problems of assuming sufficient control of the Content Information and Preservation Description Information, when they are largely digital, are addressed in four related categories as follows:

- Copyright implications
- Obtaining adequate Representation Information
- Authority to modify Representation Information
- Agreements with external organizations

Copyright implications: These issues occur when the OAIS acts as a custodian. An OAIS that acquires copyrighted material has special concerns to address. Copyright issues surrounding digital materials are in a state of flux, at the time of this writing, and an OAIS must insure that it adheres to the Copyright Law. An OAIS will have to be an active player in supporting copyright laws and thus will need to control access to some of these materials. An OAIS will need to understand the copyright concepts and applicable laws prior to accepting copyright materials into the OAIS. An OAIS armed with the knowledge and understanding of the Copyright Law can establish guidelines for ingestion of information and rules for dissemination and duplication of the information when necessary. It is beyond the scope of this document to provide details of national and international copyright laws. It is critical for an OAIS to identify any material that has a Copyright before it is ingested so that the OAIS can understand the rights it has to the material.

Obtaining Adequate Representation Information: The Representation Information for the Content Information and PDI is essential to their definition. However determining what is sufficient Representation Information is not always straight forward. For example, a scientific data set may be a file containing several types of data, including text and images and tables. The formats and meanings of each of these components, which are Representation Information, may be described in separate documents, and the overall data file structure may be described in yet another document. Each of these documents may be in digital form which means they each have additional Representation Information that is needed to understand them. Conceptually, this chaining ends when the Representation Information objects (documents) are expressed in a non-digital form that is readable and understandable to the expected Consumer's knowledge base. Practically, this chaining will often end when some of the digital Representation Information is understandable to widely available software which then provides it in a form understandable to the expected

Consumer's knowledge base. If, in the example above, the images are in the Graphics Interchange Format (GIF), the GIF Representation Information may not be included in the AIP and may only be referenced because it is assumed that GIF display software will be available to present the images.

Ideally, an OAIS would include all the digital Representation Information objects associated with a Content Information in its AIP, and it would point to the hardcopy documents needed to complete the representation chains. Further, these hardcopy documents would also be in the OAIS. When the representation chains are broken through the assumed use of widely available software, the OAIS is taking on a greater risk of information loss that needs to be managed. **These assumptions as to which software correctly presents the information and is widely available needs to be explicitly recorded in the OAIS so they can be periodically checked because eventually they will not be true.** In the above example, at some point a decision will have to be made to either include the missing GIF Representation Information, or the images will have to be migrated to a new image format and a decision made on whether this new format should be described and/or just referenced because appropriate new software is widely available.

Authority to modify Representation Information: Although the Fixity information within the Preserving Description Information of an AIP is ensuring that the Content Information related bits are not being altered, there will come a time when Content Information bits are not in a form that is convenient for the designated Consumer community. The Content Information bits may be fully documented in available hardcopy forms, so technically the information has not been lost, but practically the information has become inaccessible. The OAIS needs the authority to migrate the Content Information to new representation forms. If it is acting as a custodian, it may need to seek additional permission to make such changes. If the information is copyrighted, hopefully the OAIS already has already negotiated permission to make the changes needed to meet preservation objectives. It may need to employ subject matter experts, from outside the OAIS, to help ensure that information is not lost. Ideally, when this situation arises, both the original AIPs (fully described) and new AIPs will be retained. Migration issues are addressed more fully in Section 5.

Agreements with external organizations: An OAIS may establish a variety of agreements with other organizations to assist in its preservation objectives. It may establish an agreement with another OAIS so that it does not have to preserve all the Representation Information objects related to its Content Information objects. For example, one OAIS may hold a description of GIF that can be referenced by many other OAIS archives. This begins to form a federation of archives to achieve common purposes more effectively than could be done independently. This will work best when there is a sufficient level of trust among these archives. One way to enhance this trust is the use of archive certification programs.

Agreements with non-OAIS organizations could be useful in helping to acquire and ingest high quality SIPs that are well on their way to becoming AIPs. However an OAIS can not be responsible for information held by such organizations and such information can not be said to be under 'effective preservation for the Long Term'.

3.3 DETERMINES DESIGNATED CONSUMER COMMUNITIES

The submission, or planned submission, of Content Information and associated PDI requires a determination as to who the expected consumers of this information will be. This is necessary in order to determine if the information, as represented, will be understandable to that community.

For example, an archive may decide that certain Content Information should be understandable to the general public. It will need to be sure this information is free of jargon. For digital information, this means that the associated Representation Information must also be free of jargon and widely understandable. These concepts, and the role of software, are discussed more fully in Sections 3.3, 3.4, and 4.

For some scientific information, the designated community of consumers might be described as those with a first year graduate level education in a related scientific discipline. This is a more difficult case as it is less clear what degree of specialized scientific terminology might actually be acceptable. The producers of such specialized information are often familiar with a narrowly recognized set of jargon, so it is especially critical to clearly define the designated community for their information and to make the effort to ensure this community can understand the information.

The possible evolution of the designated community also needs consideration. Information originally intended for a narrowly defined community may need to be made more widely understandable at some future date. This is likely to mean adding explanations in support of the Representation Information and the Preserving Description Information and it can become increasingly difficult to obtain this information over time. Selecting a broader definition of the designated community when the information is first proposed for Long-term Preservation can reduce this concern and also improve the likelihood that the information will be understandable to all in the original community.

3.4 ENSURES INFORMATION IS INDEPENDENTLY UNDERSTANDABLE

The degree to which Content Information and its associated PDI conveys information to a designated community is, in general, quite subjective. Nevertheless, it is essential that an archive make this determination in order to maximize information preservation.

For example, a manuscript's Content Information may be written in English and therefore its content may be generally understandable to a wide audience. However, unless the purpose for which it was created is clearly documented, much of its meaning may be lost. This 'purpose' information is part of its Context and must be provided in the Preserving Description Information.

As another example, consider Content Information from a digital set of observations of rainfall, temperature, pressure, wind velocities, and other parameters measured all over the world for a year. This type of information is very extensive, is not usually in a form intended for direct human browsing or reading, but it is in a form appropriate to searching and

manipulation by application software. Such content may only be understandable to the original producers unless there is adequate documentation of the meaning of the various fields and their inter-relationships, and how the values relate back to the original instrumentation that made the observations. In such specialized fields extra effort is needed to ensure that the Content Information and the Preservation Description Information are understandable to a designated community. If the archive does not have this level of expertise in-house, it may need to have outside community representatives review the information for long-term understandability. Otherwise the information may be understandable to only a few specialists and be lost when they are no longer available.

Digital Content Information needs software for efficient access. However, maintaining Content Information-specific software over the long term has not yet been proven cost effective due to the narrow application of such software. The danger of information loss is great when such software is relied upon for information preservation and understanding because it may cease to function under only small changes to the hardware and software environment. This may not be recognized unless there is a vigorous, ongoing, testing program. In general, each Content Information's primary Digital Object should have its Representation Information fully documented down to the bit level, even if the lower levels of Representation Information are, at ingest time, supported by widely available software. Over time, all representations will be replaced and documenting them ensures that the OAIS can track them as to viability for usage by the OAIS Consumers. This also facilitates Content Information migration to new representations when needed.

3.5 FOLLOWS ESTABLISHED PRESERVATION POLICIES AND PROCEDURES

It is essential for an OAIS to have documented policies and procedures for preserving its AIPs, and it needs to follow those procedures. This is particularly true for digital AIPs or digital Content Information objects because of their frequent need for attention as discussed earlier. They are also highly vulnerable to inadvertent and intentional destruction or corruption even in the most trusted systems. This fragile nature puts a premium on being able to recover from all digital AIP handling operations when errors are discovered.

The appropriate policies and procedures will depend, at minimum, on the nature of the AIPs and any 'backup' relationships the archive may have with other archives. Migrations which alter any Content Information or PDI will need to be carefully monitored and the appropriate PDI fully updated. This attention to detail, while also ensuring against processing errors, requires that strong policies and procedures be in place and that they be executed.

A long-term technology usage plan, updated as technology evolves, is essential to avoid being caught with very costly system maintenance, emergency system replacements, and costly data representation transformations.

3.6 MAKES THE INFORMATION AVAILABLE

By definition, an OAIS makes its AIPs visible and available to its Consumers. Multiple views of its holdings, supported by various search aids that may cut across collections of AIPs, may be provided. Some AIPs, visible to Consumers, may not be stored as AIPs in any recognizable sense. Rather, they appear as DIPs that may be generated upon request using an associated algorithm operating on one or more existing AIPs. Upon dissemination, these DIPs will need to include documentation on how they were derived from other AIPs. The expectations of OAIS Consumers regarding access services will vary widely among archives and over time as technology evolves. Pressures for ever more effective access must be balanced with the requirements for preservation under the available resource constraints.

For security, direct access by Consumers to digitally stored AIPs should only be allowed on copies of the original AIPs. Some AIPs may have restricted access and therefore may only be disseminated to Consumers who meet access restrictions. The OAIS needs to have published policies on access and restrictions so that the rights of all parties are protected.

In general, DIPs may be distributed by all varieties of communication paths, including networks and physical media.

4 DETAILED MODELS

The purpose of this section is to provide a more detailed model view of the functional entities of the OAIS and the information handled by the OAIS. This aids OAIS designers of future systems and provides a more precise set of terms and concepts for discussion of current systems.

4.1 FUNCTIONAL MODEL

The OAIS of Figure 2-1 is separated into five functional entities and related interfaces as shown in Figure 4-1. The lines connecting entities identify communication paths over which information flows in both directions. The lines shown to Administration are dashed only to reduce diagram clutter.

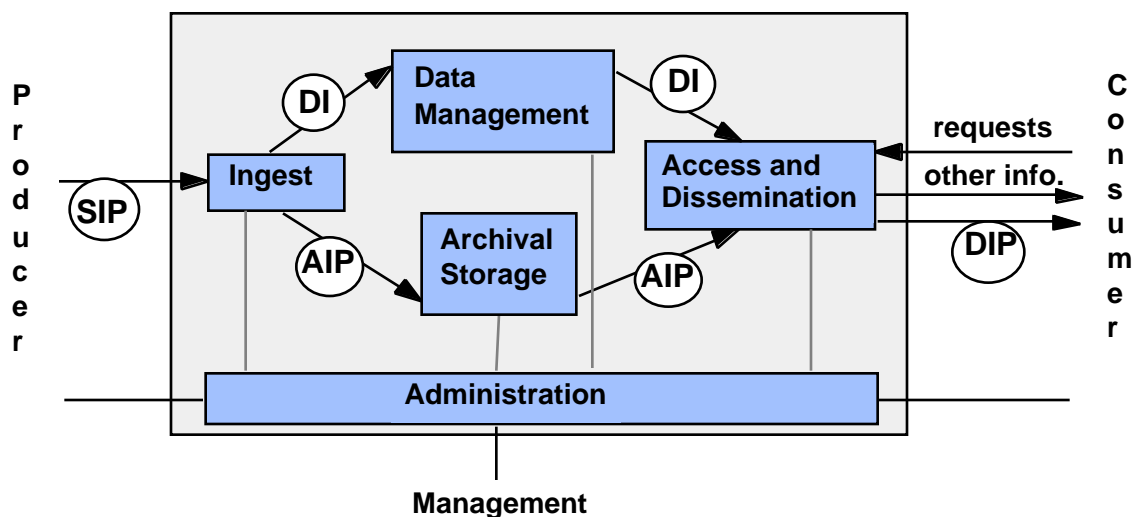


Figure 4-1. OAIS Functional Entities

The role provided by each of the entities in Figure 4-1 is briefly described as follows:

Ingest: This entity provides the services and functions to accept Submission Information Packages (SIPs) from Producers and prepare the contents for storage and management within the archive. Ingest functions include negotiating submission schedules with producers; receiving and validating submission information packages; generating an Archival Information Package (AIP) which complies with the internal archive data model and documentation standards; extracting descriptive information (DI) from the AIPs for inclusion in the archive database; and coordinating updates to Archival Storage and Data Management.

Archival Storage: This entity provides the services and functions for the storage, maintenance and retrieval of Archival Information Packages. Archival Storage functions include receiving AIPs from ingest and adding them to permanent storage; managing the storage hierarchy; refreshing the media on which archive holdings are stored; performing routine and special error checking; providing disaster recovery capabilities and providing AIPs to access and dissemination to fulfill dissemination requests.

Data Management: This entity provides the services and functions for populating, maintaining, and accessing both descriptive information which identifies and documents archive holdings and internal archive administrative data. Data Management functions include administering the archive database functions (maintaining schema and view definitions and referential integrity); performing database updates (loading new descriptive information or archive administrative data); performing queries on the data management data to generate result sets; producing reports from these result sets and activating requests triggered by time or data management updates.

Administration: This entity manages the overall operation of the archive system. The Administration functions include soliciting and negotiating submission agreements with producers; maintaining configuration management of system hardware and software and providing system engineering functions to monitor and improve archive operations; receiving and responding to status reports from all archive elements as well as feedback from consumers; developing and maintaining archive standards and policies; auditing AIPs to insure that they meet archive standards; controlling physical access to the archive system and providing a point of interaction with OAIS Management.

Access and Dissemination: This entity supports consumers in determining the existence, description, location and availability of information stored in the OAIS and allowing consumers to request and receive information products. The Access and Dissemination entity includes functions for communicating with consumers to receive requests; applying controls to limit access to specially protected information; coordinating the execution of requests to successful completion; generating Dissemination Information Packages (DIPs) by processing and combining AIPs from Archival Storage and descriptive information from Data Management; and delivering the resulting DIPs to consumers.

In addition to the entities described above, there are various common services assumed to be available, and all of these are discussed in more detail in Sections 4.1.1 through 4.1.7. Specific flows of information among the entities are shown here in bold type, followed by an identifier in braces. For example, Section 4.1.3 identifies the flow "storage confirmation {4.1.3g}." This and all other flows are then illustrated in the set of data flow diagrams collected in Section 4.1.8.

4.1.1 COMMON SERVICES

Modern, distributed computing applications assume a number of supporting services such as inter-process communication, name services, temporary storage allocation, exception handling, security, backup and directory services. Common Services provides a single conceptual source for these services.

4.1.2 INGEST

The functions of the Ingest entity are detailed below:

* The **Schedule Submission Delivery** function negotiates a data submission schedule {4.1.2k} with the producer. It also maintains a calendar of expected Data Delivery Sessions that will be needed to transfer one or more Submission Information Packages to the OAIS and the resource requirements to support their ingestion.

* The **Receive Submission** function provides the appropriate storage capability or devices to receive a SIP {4.1.2a} from the producer. Digital SIPs may be delivered via electronic transfer (e.g. ftp); loaded from media submitted to the archive; or simply mounted (e.g. CD-ROM) on the archive file system for access. Non-digital SIPs would likely be delivered by conventional shipping procedures. The Receive SIP function may represent a legal transfer of custody for the CI in the SIP, and may require that special access controls be placed on the contents. Mechanisms should be in place to validate the successful transfer of the SIP to the staging area. For digital submissions these mechanisms might include cyclic redundancy codes (CRCs) or checksums associated with each data file, or the use of system log files to record and identify any file transfer or media read/write errors. The Receive SIP function provides a confirmation of receipt {4.1.2f} of a SIP to the Producer which may include a request to resubmit a SIP in the case of errors resulting from the SIP submission.

* The **Generate Archival Information Package** function transforms one or more SIPs into one or more AIPs that conforms to the internal data model of the archive. This may involve file format conversions, data representation conversions or reorganization of the content information in the SIPs. The Generate AIP function may issue request reports {4.1.4a} to data management and to produce descriptive information {4.1.2a} needed to complete the AIP.

* The **Generate Descriptive Information** function extracts Descriptive Information from the AIPs to populate the data management system. This includes metadata to support searching and retrieving AIPs (e.g. who, what, where, when, why) and could also include special browse products (thumbnails, images) to be used as finding aids.

* The **Coordinate Updates** function is responsible for transferring the AIPs to Archival Storage and the Descriptive Information to Data Management. Transfer of the AIP {4.1.2p} includes a storage request {4.1.2h} and may represent an electronic, physical, or a virtual (i.e. data stays in place) transfer. After the transfer is completed and verified, Archival Storage returns a storage confirmation {4.1.2m} indicating the storage location of the AIP. The Coordinate Updates function also incorporates the storage location information in to the Descriptive Information {4.1.2n} for the AIP and transfers it to the Data Management entity along with a database update request {4.1.2d}. In return, Data Management provides a database update response {4.1.2e} indicating the status of the update.

4.1.3 ARCHIVAL STORAGE

The functions of the Archival Storage entity are detailed below:

* The **Receive Data** function receives a transfer request {4.1.3d} and an Archival Information Package from the staging area and moves the data to permanent storage within

the archive. The transfer request may need to indicate the anticipated frequency of utilization of the data objects comprising the Archival Information Package (AIP) to allow the appropriate storage devices or media to be selected for storing the AIP. The Transfer Receiving function will select the media type, prepare the devices or volumes and perform the physical transfer to the Archival Storage volumes. On completion of the transfer this function sends an Archival Storage confirmation {4.1.3g} message to Ingest including the storage location of the AIPs.

* The **Manage Storage Hierarchy** function positions the contents of the Archival Information Packages (AIPs) on the appropriate media based on directions from ingest (transfer request), administrative policies {4.1.3h} or usage statistics. It will also conform to any special security measures that are required and insure the appropriate level of protection for the AIP, which may dictate certain handling or backup procedures. This function provides status reports {4.1.3a} to Administration summarizing the inventory of media on-hand, available storage capacity in the various tiers of the storage hierarchy, and other operational statistics.

* The **Refresh Media** function provides the capability to reproduce the AIPs over time. Within the refresh media function the Content Information and Preservation Description Information must not be altered. However, the data constituting the Packaging Information may be changed as long as it continues to perform the same function. The migration strategy must take into consideration the expected and actual rates of errors encountered in various media types, their performance, and the costs of ownership when deciding which media to migrate to. If media dependent attributes (e.g. tape block sizes, CD-ROM volume information) have been included as part of the Content Information, a way must be found to preserve this information when migrating to higher capacity media with different storage architectures. Refer to Section 5 for a detailed description of migration issues.

* The **Error Checking** function provides statistically acceptable assurance that no components of the archive information package are corrupted during any internal archive data transfer or transformation. This function requires that all hardware and software within the archive provide notification of potential errors and that these errors are routed to standard logs that are checked by the Archival Storage staff. The PDI Fixity Information provides some assurance that the Content Information has not been altered as the AIP is moved and accessed. Similar information is needed to protect the PDI itself. A standard mechanism for tracking and verifying the validity of all data objects within the archive may also be used, for example, cyclical redundancy checks (CRCs) could be maintained for every individual data file. The storage facility procedures should provide for random verification of the integrity of data objects using CRCs or some other error checking mechanism.

* The **Disaster Recovery** function provides a mechanism for producing duplicate copies of AIPs (AIUs and AICs) in the archive collection. The backup media should be capable of being removed from the archive for storage at a separate facility. Disaster Recovery policies {4.1.3i} are specified by Administration.

* The **Provide Data** function provides copies of stored AIPs to Access and Dissemination. This function receives a data request {4.1.3j} which identifies the requested AIP(s) {4.1.3m} or potentially subsets (AIUs) or supersets (AICs) and provides them on the requested media type or transfers them to a staging area. The Provide Data function sends a notice of data transfer {4.1.3k} to Access and Dissemination upon completion of a data request.

4.1.4 DATA MANAGEMENT

The functions of the Data Management entity are detailed below:

* The **Administer Database** function is responsible for maintaining the integrity of the Data Management database which contains both descriptor information which identifies and describes the archive holdings and system information used to support archive operations. The Administer Database function is responsible for creating any schema or table definitions required to support data management functions; for providing the capability to create, maintain and access customized user views of the contents of this storage; and for providing internal validation (e.g. referential integrity) of the contents of the database.

* The **Perform Queries** function receives a query request {4.1.4b} from Access and Dissemination and executes the query to generate a result set {4.1.4x} that is transmitted to the requester.

* The **Generate Report** function receives a report request {4.1.4a} from Ingest, Access and Dissemination or Administration and executes any queries or other processes necessary to generate the report then supplies the report {4.14X} to the requester.

* The **Receive Database Updates** function adds, modifies or deletes information in the Data Management persistent storage. There are three major sources of updates; ingest transactions, system updates, and review updates. Ingest transactions consist of descriptive information which identifies new AIPs stored in the archive. System updates include all system related information (operational statistics, consumer information, request status). Review updates are generated by periodic reviewing and updating of information values (e.g. contact names, and addresses). The update function provides regular reports to administration summarizing the status of updates to the database {4.1.4e}.

* The **Activate Requests** function maintains a record of subscription requests and periodically compares it to the contents of the archive to determine if all needed data is available. If needed data is available, this function generates a Dissemination Request which is sent to the Access And Dissemination functional area. This function can also generate Dissemination Requests on a periodic basis where the length of the period is defined by the consumers or on the occurrence of an event (i.e. a database update).

4.1.5 ADMINISTRATION

The functions of the Administration entity are detailed below:

* The **Negotiate Submission Agreement** function solicits desirable archival information {4.1.5b} for inclusion into the OAIS and negotiates submission agreements with data producers {4.1.2c}. The data submission formats and procedures must be clearly documented in the archive's data submission policies and the deliverables must be identified by the producer in the submission agreement.

* The **Manage System Configuration** function provides system engineering for the archive system to systematically control changes to the configuration. This function maintains integrity and tractability of the configuration during all phases of the system life cycle. It also audits system operations, system performance, and system usage. It continuously monitors the functionality of the entire archive system and prepares recommendations and plans for system evolution. It may also provide for updates to AIPs or DI stored in Archive by providing dissemination request to Access and Dissemination, updating the contents of the resulting DIPs and resubmitting them as SIPs to the Ingest function.

* The **Physical Access Control** function provides mechanisms to restrict or allow physical access (doors, locks, guards) to elements of the archive as determined by archive policies.

* The **Develop Standards and Policies** function is responsible for developing and maintaining the archive system data standards. These standards include format standards, documentation standards and the procedures to be followed during the ingestion process. It will also develop policies {4.1.3h} for Archival Storage hierarchy management and migration policies {4.1.3h} to assure that archive storage formats do not become obsolete.

* The **Audit AIPs** function is carried out by the archive data engineers and may also involve an outside committee (e.g., science and technical review). The audit process must verify that the quality of the data meets the requirements of the archive and the review committee. It must verify that there is adequate Representation Information and Preservation Description Information to ensure the Content Information is understandable and independently usable to the Designated Community. The formality of the review will vary depending on internal archive policies. The Audit process may determine that some portions of the submission information package are not appropriate for inclusion in the archive and must be resubmitted or excluded. After the audit process is completed any liens {4.1.2b} are reported to the producer who will then resubmit {4.1.2i} or appeal {4.1.2j} the decision. After the audit is completed a final ingest report {4.1.2g} is prepared and provided to Administration and to the Producer.

* The **Interact with Management** function receives and carries out Management policies {4.1.5X}. These policies include such things as the OAIS charter, scope, resource utilization guidelines, and pricing policies. It also provides OAIS performance information {4.1.5k} to Management.

4.1.6 ACCESS AND DISSEMINATION

The functions of the Access and Dissemination entity are detailed below:

* The **Prepare Finding Aids** function provides tools and products which provide an overview of information products available in the archive system. Finding aids include special versions of products which can be quickly viewed such as thumbnails images and abstracts of documents or special interfaces (graphical selection menus) to identify and select available data. This function also generates requests for specialized queries {6a} or new dissemination requests {6b} to be carried out by the Generate DIPs function to produce new representations of the data objects to extend the retrieval capabilities of the archive (e.g., data mining).

* The **Receive Requests** function provides a single user interface to the information holdings of the archive. This interface will normally be via computer network or dial-up link to an on-line service, but might also be implemented in the form of a walk-in facility, printed catalog ordering service, or fax-back type service. Three categories of requests are distinguished, query requests {2-5C} which are executed in data management and return immediate result sets for presentation to the user; report requests {4b} which may require a number of queries and produce a formatted output for delivery to the consumer; and dissemination requests {4.1.6b} which may access either or both data management and archival storage to prepare a formal dissemination package for on- or off-line delivery. A dissemination request may be an ad-hoc request {4.1.6d} which is executed only once or a subscription request {4.1.6c} which will be maintained by the Activate Requests function in data management and may result in periodic deliveries of requested items. Other special request types are allowed but are not detailed. The Receive Request function transfers requests {request} to the Coordinate Requests function for execution.

* The **Coordinate Requests** function determines the resources needed to fulfill archive requests and forwards them to those entities for execution. Query and report requests will be fulfilled by data management while dissemination requests will generally require information from data management and archival storage. The coordinate request function will determine if resources are available to perform a request and notify the consumer that a request has been accepted (possibly with an estimate of request cost and an option to cancel the request). It will then transfer the request to data management or to the Generate DIP function for execution. The Coordinate Request function also provides assistance to archive system consumers including providing status of orders {unnumbered} and other customer support activities. The Coordinate Request function will also create, maintain and delete consumer accounts and will bill {4.1.5a} and collect payment {4.1.5c} from consumers for the utilization of archive system resources.

* The **Generate DIP** function accepts a dissemination request, retrieves the data {4.1.7m} from Archival Storage and moves a copy of the data to a staging area for further processing. This function also transmits a report request {4.1.4b} to Data Management to generate descriptive information needed to produce the DIP. If special processing is required, the Generate DIP function accesses data objects in staging storage and applies the requested processes. The types of operations which may be carried out include statistical functions, sub-sampling in temporal or spatial dimensions, conversions between different data types or output formats, and other specialized processing (e.g. image processing). This function

places the completed DIP package in the staging area and notifies both Coordinate Request and the Deliver DIP function that the package is ready for delivery.

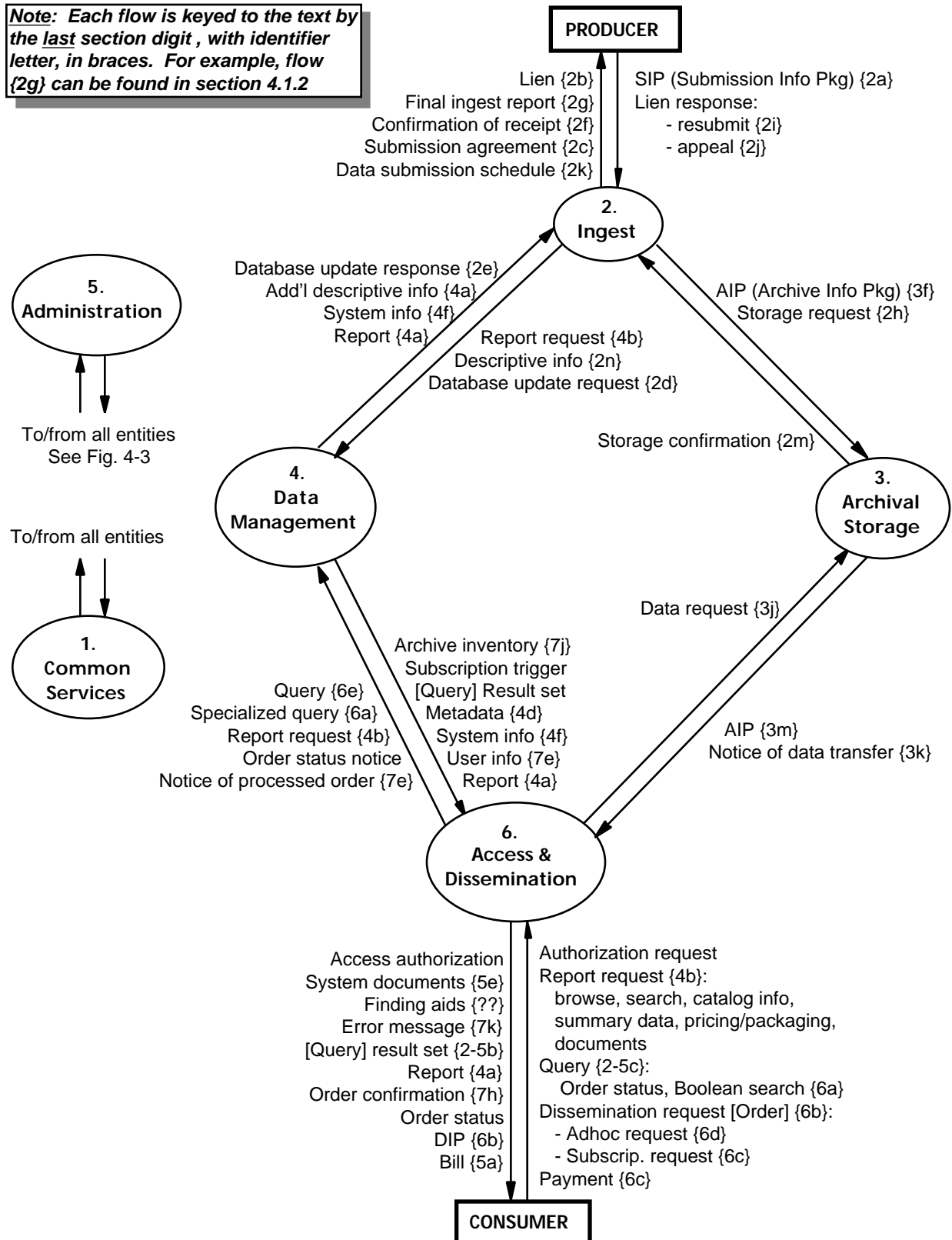
* The **Deliver DIPs** function handles both on-line and off-line deliveries of DIPs to consumers. For on-line delivery it accepts a DIP and prepares it for on-line distribution {4.1.7b} in real time via communication links. It identifies the intended recipient, the transmission procedures requested, and the DIP in the staging area to be transmitted. For off-line delivery it retrieves the DIP from the generate DIP function, prepares packing lists, bills of lading and other shipping records, and then ships the DIP {4.1.7f}. When the DIP has been shipped, a notice of processed order {4.1.7e} is returned to the Coordinate Requests functional along with billing information {4.1.7i} for completed orders.

* The **Provide Access Controls** function provides a hierarchy of security controls depending on the needs of the archive system. These include restricting access to certain information due to security classifications or copyright restrictions.

4.1.7 Data Flow Diagrams

The flow of data items among the OAIS functional entities is diagrammed in this section. Figure 4-2 shows the more significant data flows. The flows associated with the Administration are generally support background activities of the other entities. To avoid complication of Figure 4-2, these background flows are illustrated in the context diagrams of Figure 4-3.

Note: Each flow is keyed to the text by the last section digit, with identifier letter, in braces. For example, flow {2g} can be found in section 4.1.2



Paul Grunberger
Rev i, 4-6-98

Figure 4-2. OAIS Data Flow Diagram

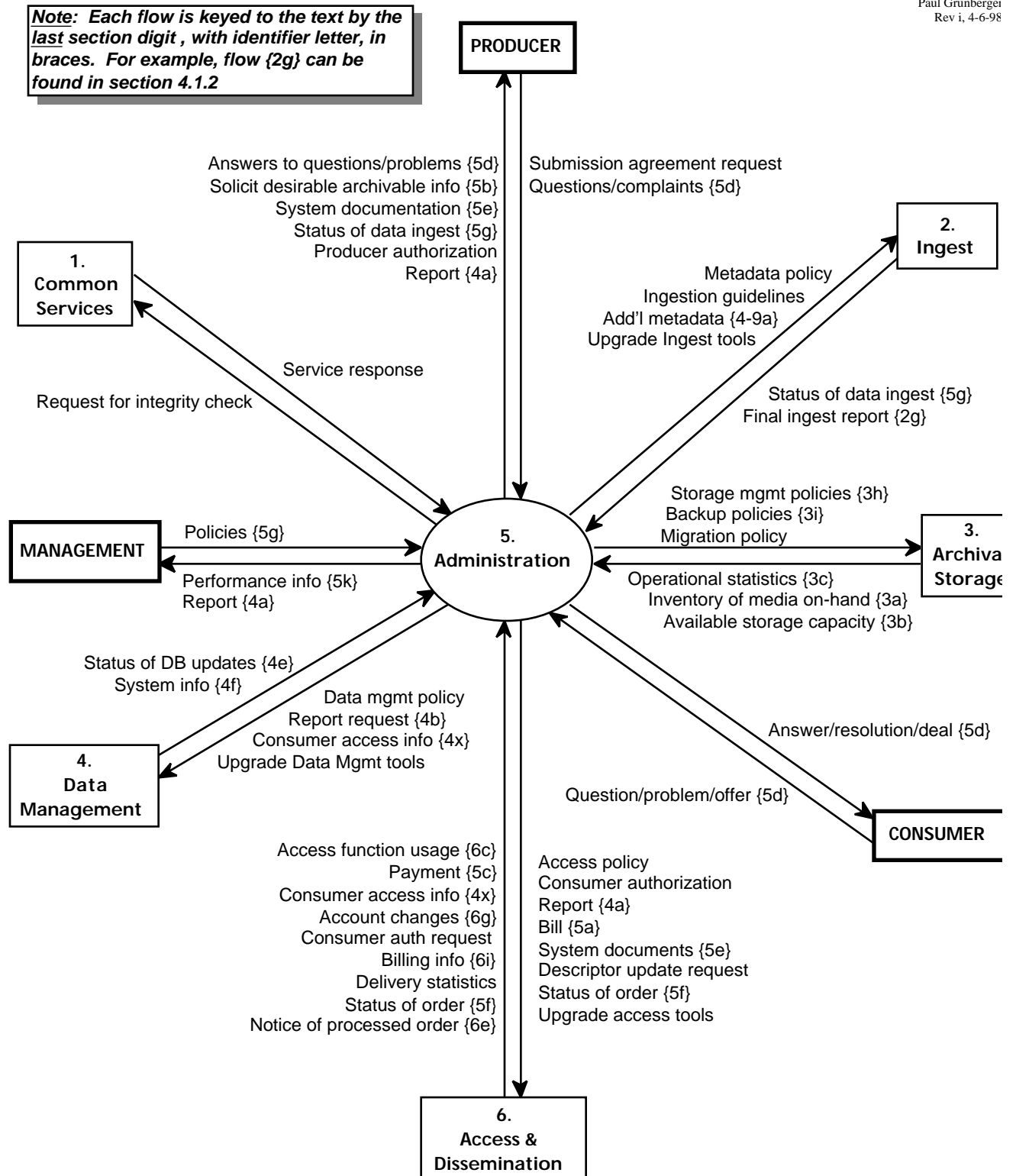


Figure 4-3. Administration Context

4.2 INFORMATION MODEL

This section builds on the concepts presented in section 2 to further describe the types of information that are exchanged and managed within the OAIS. This section also defines the specific information objects which are used within the OAIS to preserve and access the information entrusted to the archive. This more detailed model of OAIS related information objects is intended to aid the architect or designer of future OAIS systems. The objects discussed in this section are conceptual and should not be taken to imply any specific implementations.

As discussed in Section 2, the primary goal of an OAIS is to preserve information for a designated community over an indefinite period of time. In order to preserve this information an OAIS must store significantly more than the contents of the object it is expected to preserve. This section analyzes those information requirements to describe the object classes of data associated with an OAIS. This section uses “**Object Modeling Technique (OMT)** - like” [Reference 2] diagrams to model the concepts discussed in the text. An overview of the notation used and critical object modeling concepts is presented in Annex D of this document. Since the objects in this model are information structures, the dynamic modeling aspects of OMT have not been used.

Section 4.2.1 provides a model of the information required for effective long term preservation of information. Section 4.2.2 describes the conceptual objects and containers that represent the contents of an OAIS.

4.2.1 LOGICAL MODEL FOR ARCHIVAL INFORMATION

4.2.1.1 Information Object

A basic concept of the OAIS Reference Model is the concept of information being a combination of data and Representation Information. This concept is illustrated by the OMT diagram 4-5. The **Information Object** is composed of a Data Object which is either physical or digital and the Representation Information that allows for the full interpretation of the data into meaningful information. This model is valid for all the types of information in an OAIS.

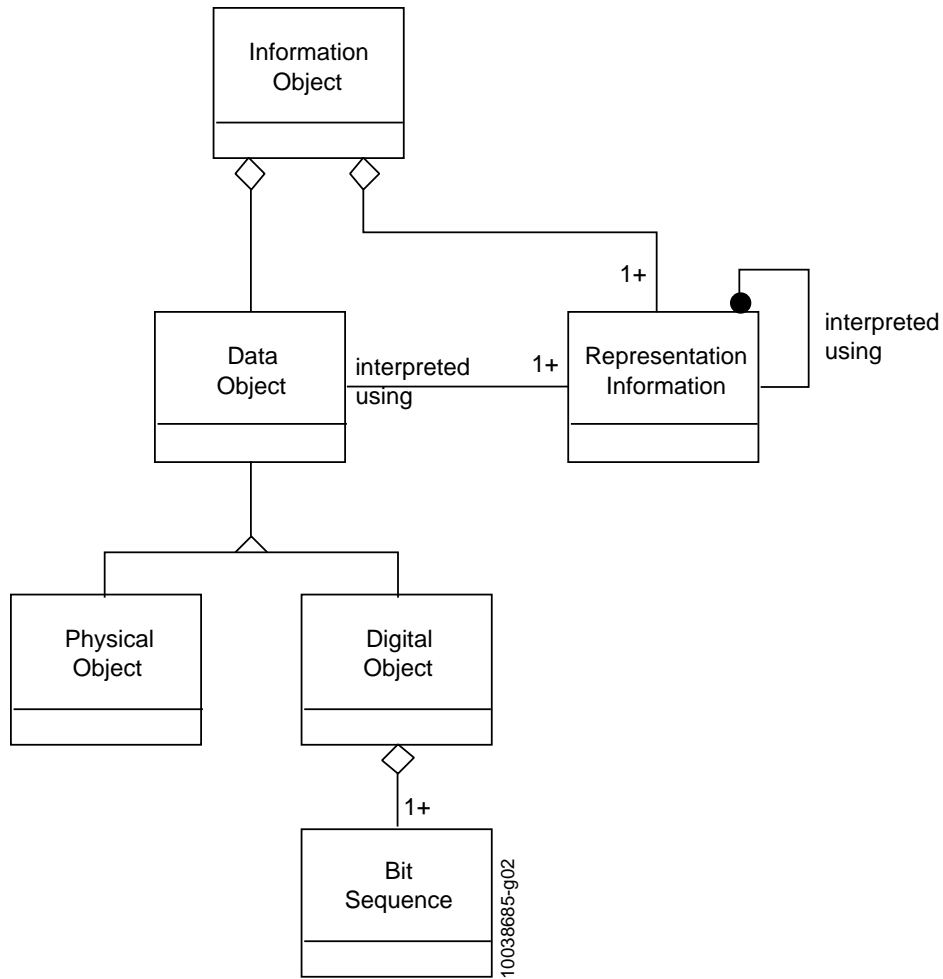


Figure 4-5. Information Object

4.2.1.2 Data Object

The Data Object may be expressed as either a physical object (e.g., a moon rock) together with some Representation Information, or it may be expressed as a digital object (i.e., a sequence of bits) together with the Representation Information giving meaning to those bits.

4.2.1.3 Representation Information

The Representation Information accompanying a physical object like a moon rock may give additional meaning, as a result of some analysis, to the physically observable attributes of the rock. This information may have been developed over time and the results, if provided, would be part of the Information Object.

The Representation Information accompanying a digital object, or sequence of bits, is used to provide additional meaning. It typically maps the bits into commonly recognized data types such as character, integer, and real and into groups of these data types. It associates these with higher level meanings which can have complex inter-relationships that are also

described.

The remainder of this section focuses on the Representation Information object when the Data Object is specialized as a Digital Object.

4.2.1.3.1 Representation Information Types

The Digital Object, as shown in Figure 4-5, is itself composed of one or more bit sequences. The purpose of the Representation Information object is to convert the bit sequences into more meaningful information. It does this by describing the format, or data structure concepts, which are to be applied to the bit sequences and that in turn result in more meaningful values such as characters, numbers, pixels, arrays, tables, etc. These common computer data types, aggregations of these data types, and mapping rules which map from the underlying datatypes to the higher level concepts needed to understand the Digital Object are referred to as the **Structure Information** of the Representation Information object. These structures are commonly identified by name or by relative position within the associated bit sequences.

The Representation Information provided by the Structure Information is seldom sufficient. Even in the case where the Digital Object is interpreted as a sequence of text characters, and described as such in the Structure Information, the additional information as to which language was being expressed should be provided. This type of additional, required information is referred to as the **Semantic Information**, although the Structure Information may provide its own set of semantics. When dealing with scientific data, for example, the information in the Semantic Information can be quite varied and complex. It will include special meanings associated with all the elements of the Structural Information, operations that may be performed on each datatype, and their inter-relationships. Figure 4-6 emphasizes the fact that Representation Information must contain both Structure Information and Semantic Information.

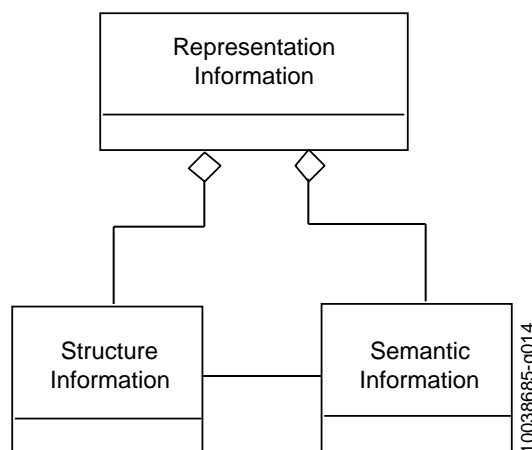


Figure 4-6. Representation Information Object

4.2.1.3.2 Representation Nets

Representation Information, which is itself an information object, may be expressed in physical forms (e.g., a paper document) or in digital forms. When the Representation Information is in digital form, additional Representation Information is needed to understand the bits of the Representation Information. In principle, this recursion continues until physical forms are encountered. For example, Representation Information expressed in ASCII needs the additional Representation Information for ASCII, which may be a physical document giving the ASCII standard. Because each Representation can be composed of multiple components, each with its own Representation, the result can be described as a **Representation Net**. In practice, the recursive chain of the representation net is also broken when there is widely available software that understands a particular representation, such as GIF display software. Nevertheless, this situation is dynamic as the technology and standards evolve and is a preservation issue needing continuing attention. Figure 4-7 illustrates the facts that Representation Information is built from:

- preexisting standards that define primitive datatypes,
- mapping rules that map those primitive datatype into the more complex datatype concept used by the Data Object,
- other semantic information that aids in the understanding of the Data such as a Data Dictionary. It also indicates the recursive nature of Representation Information as an Information Object that requires other Representation Information to enable interpretation.

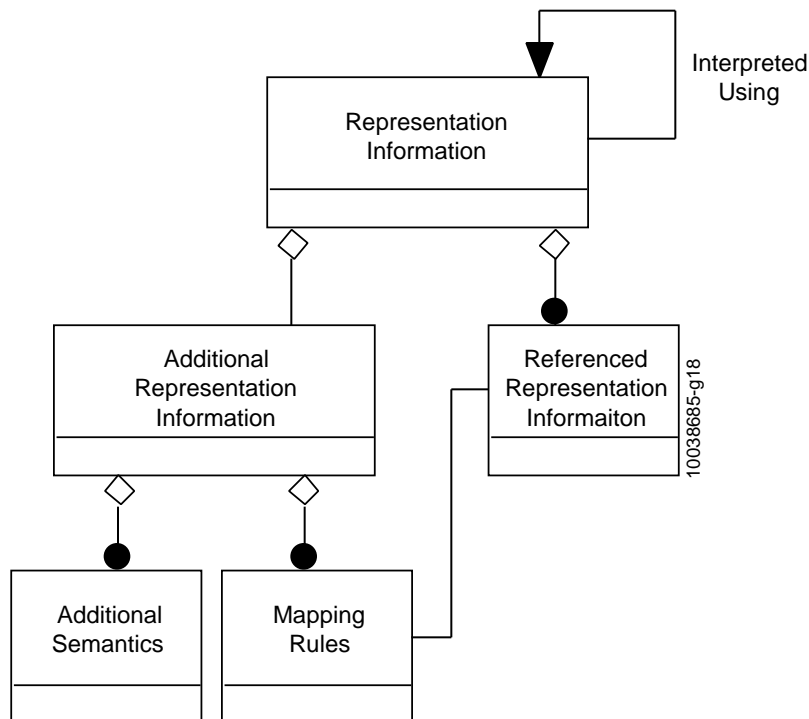


Figure 4-7: Representation Net Object Model

An example of a representation net is a multimedia document. The representation for the syntax of the content object is an international standard such as HTML for expressing on-line documents. In this hypothetical standard, the statement is made that real numbers are expressed in the ISO floating point standard, characters in ASCII and images in JPEG. In order to understand the document a consumer would need all of these underlying standards in addition to mapping rules to combine the underlying standards into the concepts required to be expressed by the document. In this example each of the underlying standards is assumed to be an ASCII file so there is no complicated Representation Net for understanding any of the representations. However, if the underlying standards were encoded in a word-processing format it would create an extension to the representation net for understanding the underlying standards.

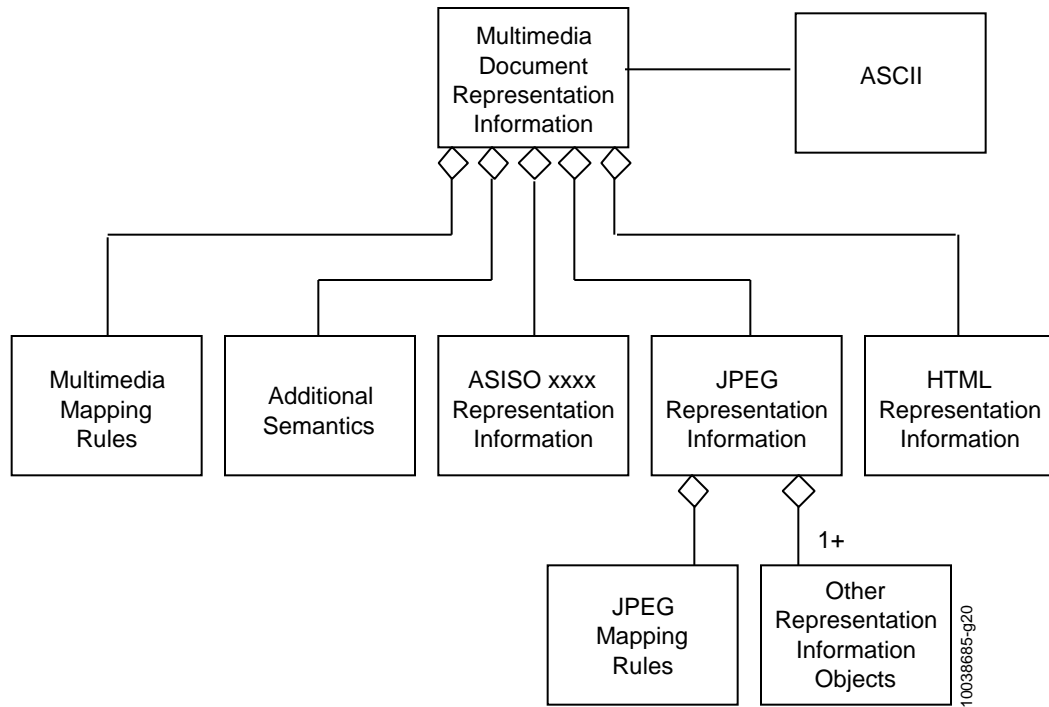


Figure 4-8 Example Representation Network

As illustrated in the previous discussion, the representation requirements for long-term preservation are much more stringent than for contemporary data processing and distribution. These issues are more fully discussed in the next section.

4.2.1.2 Taxonomy of Archival Information Object Classes

There are many types of information involved in the long-term preservation of information in an OAIS. Each of these types can be viewed as a complete Information Object in that it contains a data object and adequate Representation Information to understand the data. This section builds on the discussions in section 2.2 about the types of supporting information needed to enable long term preservation and the discussion in the previous section on the role of Representation Information. The information modeling in this section discusses several types of Information Objects that are used in the OAIS. The objects are categorized by their content and function in the operation of an OAIS into Content Information objects, Preservation Description Information objects, Packaging Information objects, And Descriptor objects. The following sections discuss the contents of each of the types of Information Object. Figure 4-9 shows a taxonomy of these Archival Information Objects.

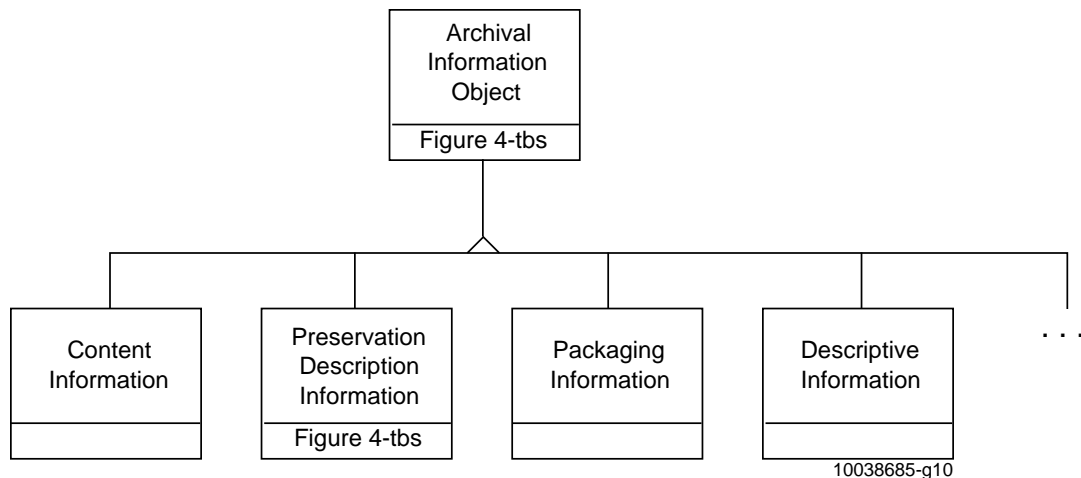


Figure 4-9. Archival Information Object Taxonomy

4.2.1.2.1 Content Information

The **Content Information** is that information which is the primary object of preservation. Deciding what is the Content Information may not be obvious and may need to be negotiated with the Producer. The Content Information can be viewed as a primary Data Object together with its Representation Information as shown in Figure 4-5 . The Data Object in the Content Information may be either a Digital Object or a Physical Object (e.g., a physical sample, microfilm). Any Information Object may serve as Content Information. The special thing about an instance of Content Information is that it is the information that an archive is tasked to preserve.

The Representation Information for the primary Digital Object (both semantic and syntactic) is needed to fully transform the bits into the Content Information. In principal, this even extends to the inclusion of definitions (e.g., dictionary and grammar) of any natural language (e.g., English) used in expressing the Content Information. Over long time periods the meaning of natural language expressions can evolve significantly in both general and in specific discipline usage. As an aside, it is clear from this example that, in general, 100% information preservation for the indefinite long-term is a goal that is not practically achievable.

The Representation Information can also be viewed as including software that supports the presentation of the Content Information to the data Consumer. Examples of this type of software include word processor's supporting complex document -format representations and scientific visualization systems supporting representations for expressing time series or multidimensional arrays. The software uses its knowledge of the underlying Representation Information to provide these services. This is useful long as the software executes properly. However for indefinite long-term information preservation, a full and understandable

description of the Representation Information is essential. This can be a particular challenge for the preservation of scientific type data as there are few standards for how to express the full range of scientific information and archives need to ensure this information is understandable to the designated Consumer communities.

As a practical matter, the OAIS needs to have enough Representation Information associated with the bits of the Data Object in the Content Information that it feels confident that it, and those expected to use the Content Information, can enter the Representation Net with enough knowledge to begin accurately interpreting the Representation Information. This is a significant risk area for an OAIS, particularly for those with a narrow discipline focus, because jargon and apparently widely understood terms may subsequently be found to be quite temporary. In such cases extra care needs to be exercised to ensure that the natural evolution of what is commonly understood in those communities does not effectively cause information loss from the Content Information .

Often required information will be embedded in the software packages used by the narrow community to present and analyze the Content Information. Unless this information is kept in other forms, this is a dangerous practice because as technology evolves many software products will become incompatible with new software/hardware platforms and cease to function. Extracting such information from non-working software can be very difficult.

An important function of the OAIS is deciding which parts of the Content Information are the primary Digital Object and what parts are the Representation Information. This aspect is critical to a clear understanding of what is being preserved .

The identification of the Content Information with its Representation Information objects can be addressed by a series of steps, as follows:

1. Identify the bits comprising the primary Digital Object of the Content Information
2. Identify one or more Representation Information objects that together address all the bits and convert them into more meaningful information.
3. For each Representation Information object identified in Step 2, examine each to identify if it is an Importing Representation Information object. If it is, identify the Representation Information objects it incorporates by reference. Repeat this step until no additional Representation Information objects are identified.
4. For each new Representation Information object identified in Steps 2 and 3 above, that is held as a Digital Object, repeat Steps 2 through 4 until no new Representation Objects are identified.
5. The Content Information consists of the primary Digital Object and each of the Representation Information objects identified in Steps 2 through 4.

As an example of this practice, consider an electronic file containing a sequence of values

obtained from a sensor looking at the Earth's environment. There is a second file, encoded using ASCII, that provides information on how to understand the first file. It describes how to interpret the bits of the first file to obtain meaningful numbers, it describes what these numbers mean in terms of the physics of the observation being conducted, it gives the date and time period over which the observations were made, it gives an average value for the observed values, and it describes who made the observations. These two files are submitted to an OAIS for preservation.

Assume that the OAIS determines that the Content Information to be preserved is the observed bits together with their values as numbers and the physics meaning of these numbers. This information is conveyed by the bit sequence within the first file together with the **Representation Information** from the second file needed to transform the first file's bits into meaningful physical values. Note that neither the first file's underlying media nor the particular file system carrying the bits is part of the Content Information in this example. Only part of the second file's content is considered a part of the Content Information and this is the part that enables the transformation of the bits from the first file into meaningful physical values. In fact this second file does not carry all the Representation Information needed to make this transformation because the following additional information is needed:

- Information that the second file is encoded in ASCII so that it can be read as meaningful characters;
- Information on how the characters are used to express the transformations from bits to numbers to meaningful physics values. This information, typically referred to as a combination of format information and data dictionary information, may also include instrument calibration values and information on how the calibrations are to be applied. All this information may be widely understandable once the ASCII characters are visible because it has all been expressed in English (or some other natural language), or some of it may be in more structured forms that will need additional Representation Information to be understood.

Therefore the Representation Information of the second file needs additional Representation Information, and this information may need additional Representation Information, etc., forming a linked set of Representations of Representations. This is a good example of the complex Representation Net.

Recall that in the example above, there was a determination that the Content Information consisted of the observed sensor values and their meanings. This is by no means the only determination that could have been made. It could just as easily have been determined that the primary Digital Object of the desired Content Information was the bit sequences within the first file together with all the bit sequences within the second file. The fact that some of these latter bit sequences are used to interpret the first file's bit sequences is just an example of a set of bits that is somewhat self-describing. It is irrelevant that some of the bits in the second file are the basis for information on the date and time period over which the observations were made, the average value for the observed values, and who made the observations. Once it has been determined that all these bits constitute the primary Digital Object of the Content Information, then the Representation Information is that information

needed to turn them into meaningful information. How extensive this meaning is to be carried and how far the Representation Net needs to be carried are local issues for the OAIS and its related Producer and Consumer communities.

4.2.1.2.2 Preservation Description Information

In addition to Content Information, the Archival Information must include information which will allow the understanding of the Content Information over an indefinite period of time. The specific set of Information Objects which are required for this function are called collectively called **Preservation Description Information** (PDI). The PDI must include information which is necessary to adequately preserve the particular Content Information with which it is associated. It is specifically focused on describing the past and present states of the Content Information, ensuring it is uniquely identifiable, and ensuring it has not been unknowingly altered.

This information is typical for all types of archives and has been classified in the context of traditional archives. However, the class definitions must be extended for digital archives. The following definitions are based on the categories discussed in the paper “Preserving Digital Information” (REF. 2) However the “Preserving Digital Information” information model did not include the separate information object classes for the Packaging Information and Descriptor Information which are included in this OAIS Information Model. Therefore the following PDI class definitions are subsets of those discussed in that paper with some of the information allocated to the new Packaging and Descriptor Objects. The differences between the OAIS information model and the information model presented in the “Preserving Digital Information” are detailed below. Table 4-1 gives examples of this information for various popular Content Information types.

- **Reference Information:** This information identifies, and if necessary describes, one or more mechanisms used to provide assigned identifiers for the Content Information. It also provides those identifiers that allow outside systems to refer, unambiguously, to this particular Content Information. In the OAIS Reference Model most if not all of this information is replicated in Package Descriptors which enable Consumers to access Content information of interest.
- **Context Information:** This information documents the relationships of the Content Information to its environment. This includes why the Content Information was created, and how it relates to other Content Information objects existing elsewhere. The Context Information differs from the PDI definition in that it does not include the information used in associating logical information with physical media. This type of information is assigned to the Packaging Information in the OAIS Reference Model.
- **Provenance Information:** This information documents the history of the Content Information. This tells the origin or source of the Content Information, any changes that may have taken place since it was originated, and who has had custody of it since it was originated. This gives future users some assurance as to the likely reliability of the Content Information. Provenance can be viewed as a special type of context information.

- **Fixity Information:** This information documents the authentication mechanisms, and it provides any authentication keys used to ensure that the particular Content Information object has not been altered in an undocumented manner. Fixity mechanisms are distributed among the security mechanisms provided by the OAIS underlying services, error protection supplied by the media and device drivers used by Archival Storage, and special encoding and error correction schemes that may be specific to a single Content Object or a family of Content Objects. The Fixity Information may include descriptions of each of these mechanisms but must at a minimum contain a description of any Content Object specific Fixity mechanisms

Content Information Type	Reference	Provenance	Context	Fixity
Space Science Data	Object Identifier Journal Reference Mission, instrument, and title attribute set	Instrument Description Processing History Sensor Description Instrument Instrument mode Processing history Decommuration map Software Interface Spec.	Calibration history Related data sets Mission Funding history	CRC Checksum Reed-Solomon coding
Bibliographic Information	ISBN Title Author	Printing history Copyright Position in series Manuscripts References	Related References Dewy Decimal System Publishing Date Publisher	Author Digital signature Cover
Software Package	Name Author Version number Serial Number	Revision History License holder Registration Copyright	Help file User Guide Related Software Language	Certificate Checksum Encryption CRC

Table 4-1: Examples of PDI Types

The OAIS needs to explicitly decide what the exact definition of Content Information is in order to be able to ensure that it also has the PDI needed to preserve the Content Information. Once the Content Information has been determined, it is possible to assess the Preservation Description Information.

4.2.1.2.3 Packaging Information

The **Packaging Information** is that information which, either actually or logically, binds and relates the components of the package into an identifiable entity on specific media. . For example, if the Content Information and PDI are identified as being the content of specific

files on a CD-ROM, then the Packaging Information may include the ISO-9660 volume/file structure on the CD-ROM. The Packaging Information, in this case, may also include the physical CD-ROM disk. These choices are the subject of local archive definitions or conventions. The Packaging Information does not necessarily need to be preserved by an OAIS since it does not contribute to the Content Information or the PDI.. However there are cases where the OAIS may be required to reproduce the original submission exactly, in which case the Content Information is defined to include all the bits submitted.

The OAIS should also avoid holding PDI or Content Information only in the naming conventions of directory or file name structures. Since these structures are most likely to be used as Packaging Information and Packaging Information is not preserved this information may be lost when the Packaging Information is altered.. The subject of Packaging Information is an important consideration to the Migration of Information in an OAIS to newer media. This subject is addressed in detail in section 5 of this document.

4.2.1.2.4 Descriptive Information

The Information Objects described previously in this section provide the information necessary to enable the long term preservation function of the archive. In addition to preserving information, the OAIS must provide adequate features to allow consumers to locate information of potential interest, analyze that information, and order desired information. This is accomplished through a specialization of the Information Object called Descriptive Information which contain the data that serves as the input to documents or applications called **Access Aids**. Access Aids can be used to locate, analyze, retrieve, or order information from the OAIS. The information needed for one access aid is called an **Associated Description**. A single object of Descriptive Information may contain several Associated Descriptions depending on the number of different access aids that can locate, visualize, retrieve or order the associated Content Information and PDI .

4.2.2 LOGICAL MODEL OF INFORMATION IN AN OPEN ARCHIVAL INFORMATION SYSTEM (OAIS)

The previous section defines the types of Information Objects that are needed by an OAIS to enable the long term preservation of information and effective access to the preserved information by the designated community. This section uses those Information Objects descriptions to model the conceptual information structures required to accomplish these functions. The models presented in this section are not intended to imply a concrete implementation but rather to highlight the relationship among the types of information needed in the archival process.

4.2.2.1 Information Package

The conceptual structure for supporting long term preservation of information is the Information Package. The OMT diagram in figure 4-10 illustrate the conceptual view of an Information Package. An Information Package is a container which contains two types of information objects, the Content Information and the Preservation Description Information

(PDI); The information package can be associated with two other types of information objects, Packaging Information and Package Descriptors. The Information Package is used in several contexts within the archival process.

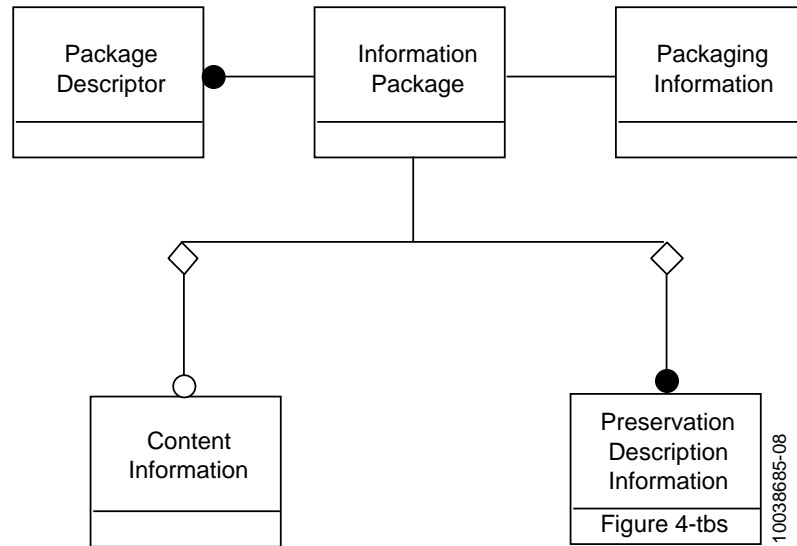


Figure 4-10. Information Package Contents

4.2.2.2 Types of Information Packages

There are three subtypes of the Information Package identified in section 2.2, the Submission Information Package (SIP), the Archival Information Package (AIP), and the Dissemination Information Package (DIP). The definitions of these package types in section 2 is based on the function of the archival process which uses the package and the translation from one package to another as it passes through the archival process. This taxonomy of Information Package types is shown in Figure 4-11.

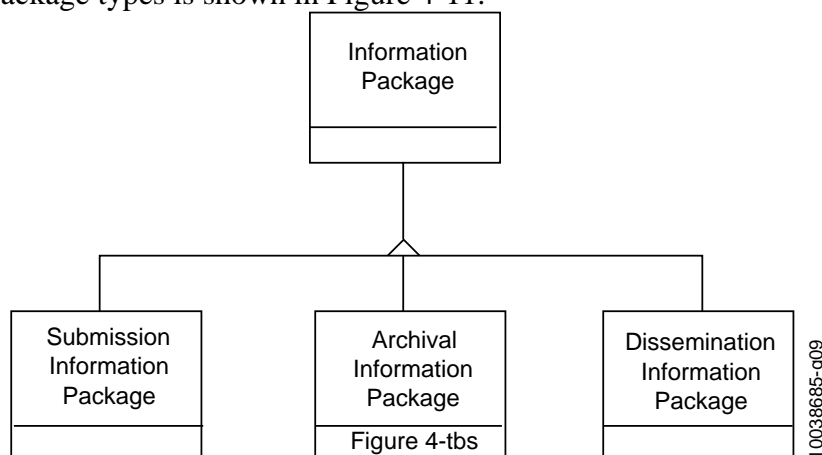


Figure 4-11. Information Package Taxonomy

The **Submission Information Package (SIP)** is that package that is sent to an OAIS by a Producer. Its form and detailed content is typically negotiated between the Producer and the OAIS. Most SIPs will have some Content Information and some PDI, but it may require several SIPs to provide a complete set of Content Information and associated PDI to form an AIP. The Packaging Information will always be present in some form.

Within the OAIS one or more SIPs is transformed into one or more **Archival Information Packages (AIP)** for preservation. The AIP has a complete set of PDI for the associated Content Information. The AIP may also contain a collection of other AIPs and this is discussed and modeled in Section 4. The Packaging Information of the AIP will conform to OAIS internal standards, and it may vary as it is managed by the OAIS.

In response to a request, the OAIS provides all or a part of an AIP to a Consumer in the form of a **Dissemination Information Package (DIP)**. The DIP may also include collections of AIPs, and it may or may not have complete PDI. The Packaging Information will always be present in some form so that the Consumer can clearly distinguish the information requested. The Packaging Information may take several forms depending on the dissemination media and Consumer requirements.

Though the implementation of the AIP may vary from archive to archive, the specification of the AIP as a container that contains all the needed information to allow long-term preservation and access to archive holdings remains valid. The information model for the AIP presented in the section 4.2.2.3. should be used as a reference to establish the types of information required to enable long-term preservation and access.

The exact information contents of the SIP and DIP and their relationship to the corresponding AIP are dependent on the agreements between the archive and its Producers and Consumers. The model for both of these packages is the same as for the Information Package with the same cardinality and optionality among contained Information Objects. The subject of transformations between SIP and AIP, and between AIP and DIP is further discussed in Section 4.3.

4.2.2.3 The Archival Information Package

An **Archival Information Package (AIP)** which is modeled in Figure 4-12 is a specialization of the Information Package which is defined to provide a concise way of referring to a set of information that has, in principle, all the qualities needed for permanent, or indefinite-long term, preservation of a designated Information Object. The AIP is itself an information object that is a container of other information objects. Within the AIP is the designated information object and it is called the Content Information.

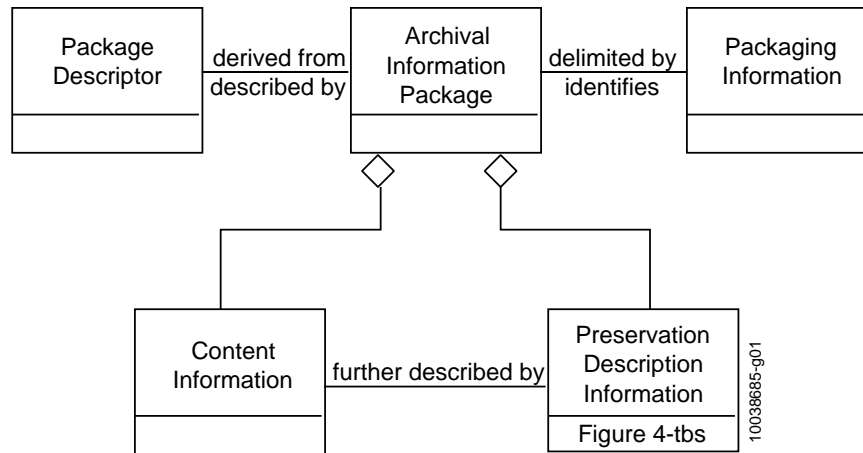


Figure 4-12. Archival Information Package (AIP)

Also within the AIP is an information object called the **Preservation Description Information (PDI)**. The PDI contains additional information about the Content Information and is needed to make the Content Information meaningful for the indefinite long-term.

The Preservation Description Information requirements in an AIP are much more stringent than the requirements for Preservation Description Information in the general Information Package. While no PDI objects are mandatory in an Information Package, all classes of PDI information must be present in an AIP. This is illustrated in Figure 4-13.

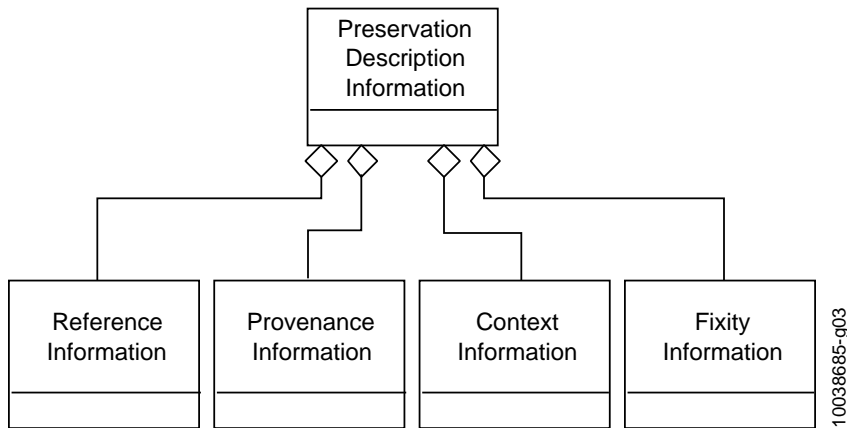


Figure 4-13. PDI Preservation Description Information

The AIP is delimited and identified by the **Packaging Information**. The **Packaging Information** may actually be present as a structure on the media that contains the AIP or may be virtual in that it is contained in the OAIS Archival Storage function. However, the delimitation and internal identification functions must be well defined in an OAIS.

Each AIP is associated with a structured form of Descriptive Information called the **Package Descriptor** which enables the consumer to locate information of potential interest, analyze that information, and order desired information.

The Package Descriptor must contain one Associated Description which allows authorized users to retrieve the Content and PDI described by the Package Descriptor. In most current archives, only internal archive processes and operations personnel and functions are authorized to use this Access Aid. However, as technology advances increase the processing power of the archive and the bandwidth between the archive and the user, such access methods as “content based queries” and “data mining” may provide the user with direct read only access to the content Information.

The Package Descriptor is not required for the long-term preservation of the Content Information but is needed to provide visibility and access into the contents of an archive. The contents of the Package Descriptor are highly dependent on the structure of the Content Information and PDI it describes. The uses and types of product descriptors in an OAIS are further defined in the section 4.2..3.1.

Figure 4-14 gives a detailed view of the Archival Information Package by expanding the PDI and the Content Information. All the "contains " relationships discussed in this section are logical containment relationships. This type of containment relationship may be physical or may be accomplished via a pointer to another object in storage.

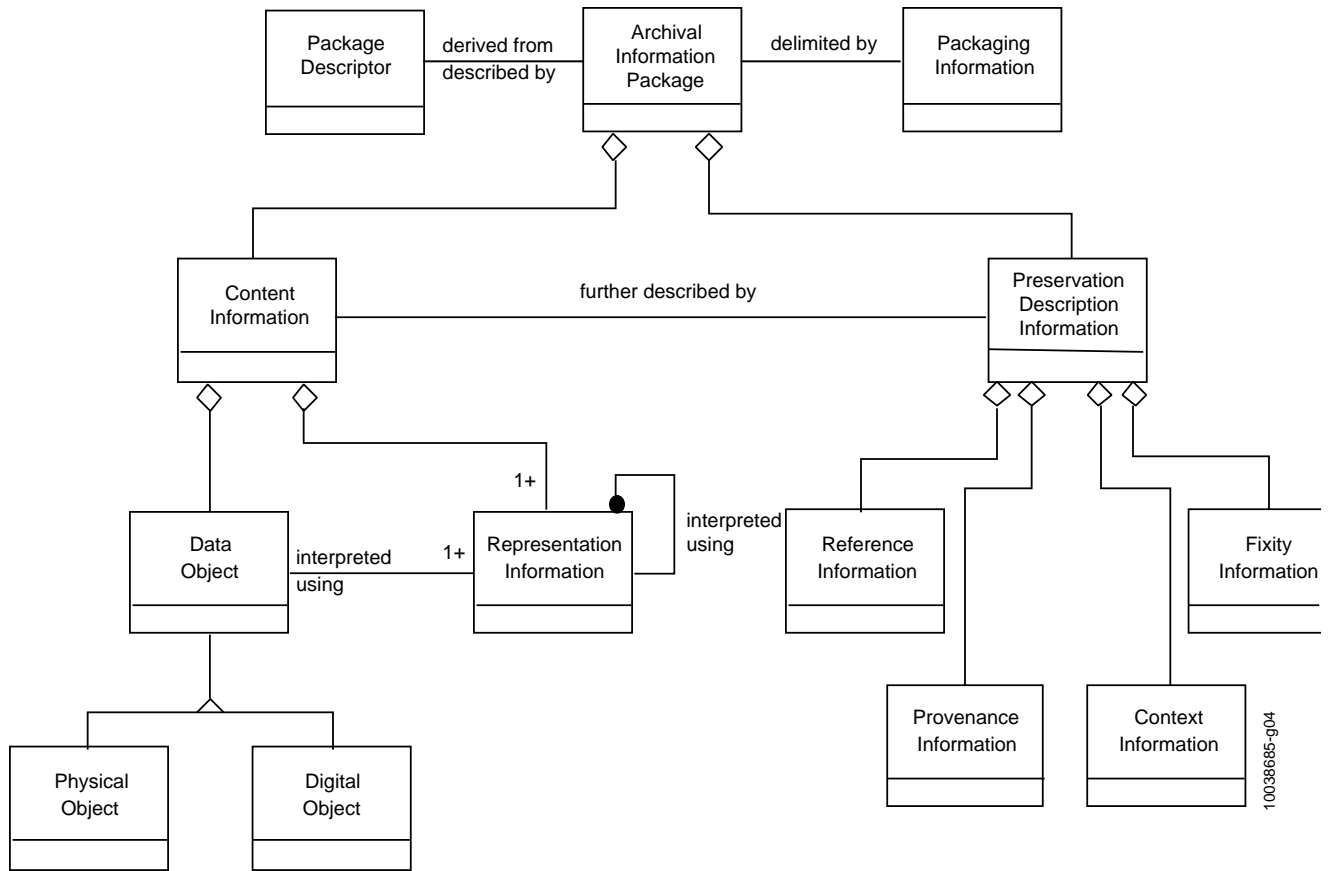


Figure 4 -14. Archival Information Package (Detailed View)

4.2.3.1. Specialization of the AIP and Associated Descriptors

Two important specializations of the AIP are discussed in this section, the **Archival Information Unit (AIU)** and the **Archive Information Collection (AIC)**. 4-15 is an OMT diagram illustrating this specialization. Both AIU and AIC are subtypes of the AIP and as such contain constructs to enable both long term preservation and consumer access. The AIU represents the type used for the preservation function of a single content atomic object. The AIC organizes a set of AIPs (AIUs and other AIPs) along a thematic hierarchy which can support flexible and efficient access by the consumer community. Conceptually all the AIPs organized by an AIC are contained in the content field of that AIC. The differences between AIUs and AICs is the complexity of their Content Information and their associated Package Descriptors and Packaging Information.

This reference model considers the differences in the Content Information and associated Packaging and Descriptor functionality between AIU and AIC to be adequately complex and linked to justify the definition of separate object classes.

From an Access viewpoint, new subsetting and manipulation capabilities are beginning to blur the distinction between AICs and AIUs. Content objects which used to be viewed as atomic, can now viewed as containing a large variation of contents based on the subsetting

parameters chosen. In a more extreme example, the Content Information of an AIU may not exist as a physical entity. The Content Information could consist of several input files (or pointers to the AIUs containing these data files) and an algorithm which uses these files to create the data object of interest.

From an information preservation viewpoint the distinction between AIU and AIC remains clear. An AIU is viewed as having a single content information object which is described by exactly one set of PDI. An AIC content object is viewed as a collection of other AICs and AIUs, each of which has its own PDI. In addition, the AIC has its own PDI which describes the collection criteria and process.

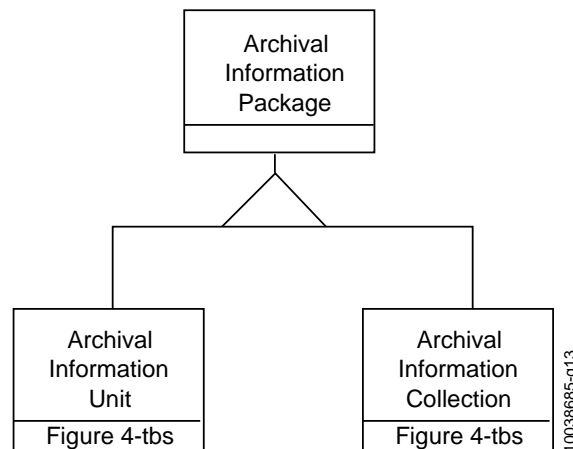


Figure 4-15. Archival specialization of the AIP

There are two specializations of the Package Descriptor, the Unit Descriptor and the Collection Descriptor. Figure 4-16 is an OMT diagram illustrates this specialization. The difference in these two classes is based on the functionality needed to effectively access the contents of an atomic AIU versus the functionality needed to effectively access the contents of an AIC.

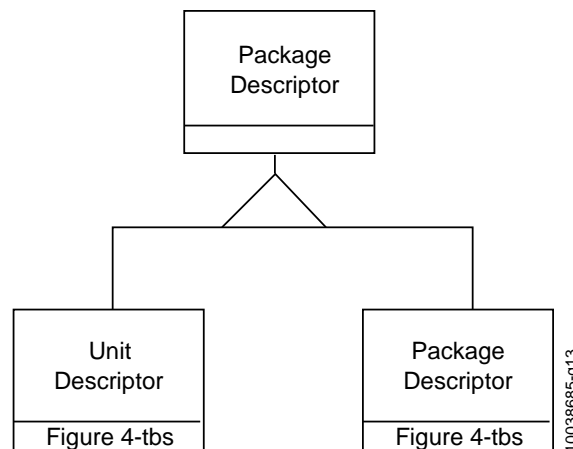


Figure 4-16. Archival specialization of the Package Descriptor

To aid in the understanding of these constructs the next two sections of this document will use an example of a company setting up an OAIS of digital versions of films. This example will focus on the information content of constructs in an AIP. Section 4.3 will illustrate more of the details of the information transformations and dataflows in an OAIS.

4.2.3.1.1. Archive Information Unit

The AIUs can be viewed as the "atoms" of information that the archive is tasked to store. A single AIU contains exactly one Content Information object (which may consist of multiple files) and exactly one set of PDI. When an information object is ingested into the OAIS a **Unit Descriptor**, which is a subtype of a Package Descriptor is created by extracting information from the Content Information and the PDI and adding OAIS specific information such as a unique identifier. The AIU is illustrated in Figure 4-17.

In the example of a digital film OAIS, the AIU for a single film can be viewed as three objects, one containing a digital image of the film in a proprietary format, one containing the Representation Information needed to understand the proprietary format (these two objects form the Content Information), and the other containing facts about the film such as date of creation, featured actors, director, producer, sequels, movie studio, and a checksum to ensure the integrity of the digital image (PDI). Since the OAIS reference model is implementation independent, these objects could be implemented as one file or multiple files. This type of implementation dependent information is contained in the Packaging Information. When a movie is ingested into the OAIS a Unit Descriptor is created by extracting information from the Content Information and the PDI and appending it to the unique ordering information..

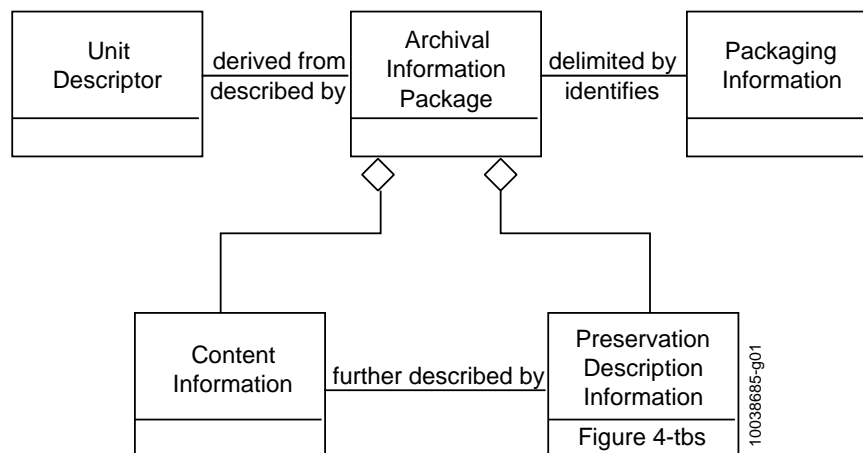


Figure 4-17. Archival Information Unit (AIU)

4.2.3.1.2 Unit Descriptor

The Unit Descriptor is a specialization of the Package Descriptor that always contains:

a set of **Associated Descriptions** each of which describe the AIU Content Information from the point of view of a single **Access Aid**. Figure 4-18 is an OMT diagram that illustrates the Unit Descriptor contents.

An important type of the Associated Description in the Unit Descriptor supplies data for a **Retrieval Aid** which enables authorized users to retrieve the AIU described by the Unit Descriptor from Archival Storage. In most current archives, only internal archive processes and operations personnel are authorized to use these Access Methods. However, as technology advances increases the processing power of the archive and the bandwidth between the archive and the user such access methods as “content based queries” and “data mining” may provide the user with direct read only access to the content Information

For example, a pattern recognition technique might be created for digital movies and the digital movie OAIS might offer a service to search its archives for large structures such as the pyramids or a New York skyline. Note that this type of service is very processing intensive, involving potentially large numbers of AIUs to be transferred from Archival Storage to Dissemination and then running the appropriate process to analyze the Content Information from each AIU. If the results are generally useful, the archivist could summarize the results of this “content based query” into a new Associated Description. This technique is frequently referred to as data mining.

An important type of Access Aid is the **Finding Aid** which is an application that assists the consumer in locating information of interest. A single AIU may have a number of Associated Descriptions which describe the Content Information using different technologies. Additionally as new description extraction and display technologies become available an archive may want to update the Unit Descriptor associated with each of its AIUs to add a new Associated Description that utilizes the new technology to better describe the AIUs.

In the digital movie OAIS example, initially, there may be one Associated Description that is a free text description of a movie, another that is a five minute clip and another that is a row in a relational database that is used by film collectors to locate films of interest. After the archive has been operational for a period of time a technique for supplying compressed digital movies may be developed based on recording every tenth frame. The archivist may decide to create an additional type of Associated Description which is populated using the results of this new technique. If necessary, the user can run each of the AIUs contained in the archive through this compression technique and create a new Associated Description for each movie in the archive or simply include this Associated Description for new movies ingested into the OAIS.

Another important class of Associated Descriptions supply data for **Ordering Aids** that allow the consumer to discover the cost of and order AIUs of interest. The Ordering Aids also allow users to specify transformations to be applied to the AIUs prior to dissemination. These transformations can include data object transformations such as subsetting, subsampling or format transformations. The transformations can also involve modifying the PDI in the AIU prior to dissemination.

For example, the digital movie OAIS could allow a user to order a digital movie as a VHS tape, a laser disc or an MPEG object delivered on-line. Each of these would involve a format transformation and in theory an update to the PDI information in the AIP to create accurate PDI for the DIP.

Although, the AIU and its associated Unit Descriptor provide the information necessary for a consumer to locate and order AIUs of interest, it can be impossible for a consumer to sort through the millions of Unit Descriptors in a large archive. This problem is addressed in the OAIS through the Archive Information Collection discussed in the next section.

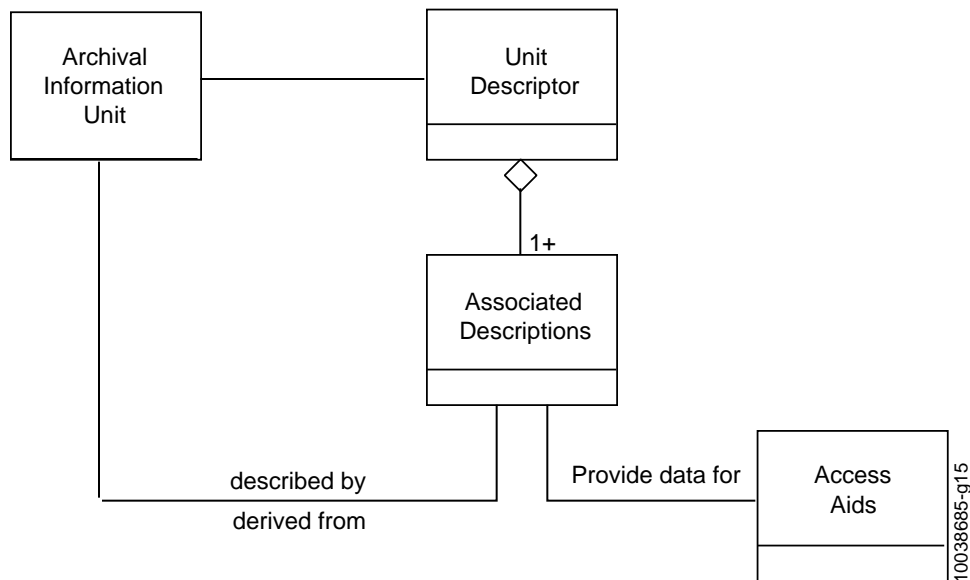


Figure 4-18 Unit Descriptor

4.2.3.1.3 Archive Information Collections

The content information of an AIC is composed of complete AIPs each of which have their own Content Information, PDI, and associated Packaging Information and Package Descriptors. These AIPs are then aggregated into Archive Information Collections (AIC) using criteria determined by the archivist. Generally AICs are based on the AIUs of interest having common themes or origins and a common set of Associate Descriptions. At a minimum all OAIS can be viewed as having at least one AIC which contains all the AIPs held by the OAIS.

A logical model of a AIC is shown in Figure 4-19. As in the previous sections all the containment relationships are logical containment and may be physical or may be accomplished via a pointer to another object in storage. For example, the Content

Information of an AIC can be created either by creating physical collections of the contained AIPs or by pointing to the contained AIPs. A single AIP can belong to any number of AICs.

For example the digital movies OAIS may have AICs based on the subject area of the movie such as mystery, science fiction , or horror. In addition the archive may have AICs based on other factors such as director or lead actor.

An important feature of the AIC as shown in Figure 4-19 is the fact that an AIC is a complete AIP which contains PDI which provides further information about the AIC such as provenance on when and why it was created, context to related AICs, and Fixity information. This is in addition to the PDI contained in member AIPs. This type of information is often necessary for a Consumer to have confidence in the reliability of an AIC. In the digital movies OAIS example, the usefulness of an AIC of movies starring John Wayne is to some extent based on the provenance of when the collection was created or last updated.

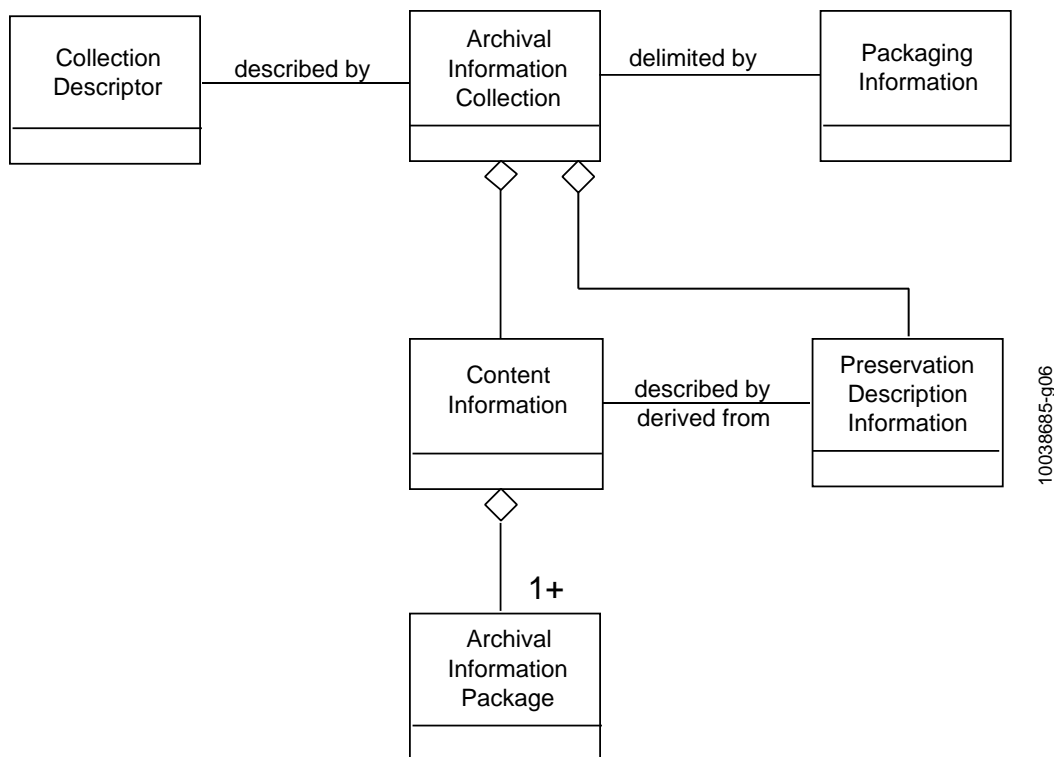


Figure 4-19. Archive Information Collections Logical View

4.2.3.1.4 Collection Descriptors

The **Collection Descriptor** is a subtype of the Package Descriptor which has added structures to better handle the complex content information of an AIC. The Collection Descriptor, which is modeled in Figure 4-20 contains the information classes which are contained in the Unit Descriptor. The Retrieval Aid of a Collection Descriptor provide a user

with access to the entire set of Content Information of the associated AIC and the PDI for the AIC, but not necessarily the individual AIPs contained in of the AIC.

There are two types of Associated Descriptions in an Collection Descriptor:

- Associated Description that describe the collection as a whole (called Collection Description in Figure 4-20)
- Associated Description that separately describe each member of the collection (called Member Description in Figure 4-20)

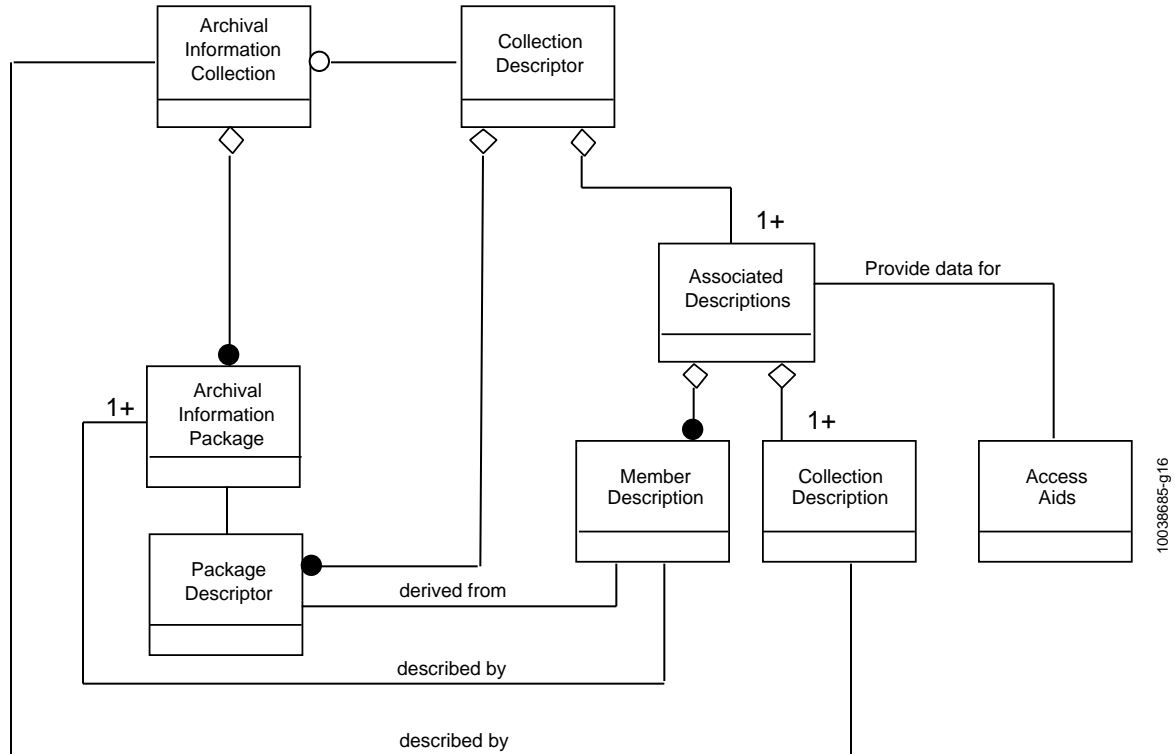


Figure 4-20. Collection Descriptors

In addition to the information contained in a Unit Descriptor, the Collection Descriptor optionally contains the Package Descriptors of the AIPs contained in the AIC. This containment relationship is logical in that the AIC may either include the Package Descriptors of member Information Packages directly or, more commonly, use pointers to the Package Descriptors of the member Information Packages. This list of the Package Descriptors for contained AIPs in an AIC could provide Access Aids with a method to access or order individual members of the AIC. It also allows alternative concepts for the implementation of Member Descriptions. They could be implemented either in a centralized fashion as an Associated Description in the Collection Descriptor or in a distributed fashion by searching the Associated Description of each member Package Description.

Another important benefit of the Collection Descriptors is the ability to define new **Theme Collections** which are not organized around existing AICs. A Theme Collection may be based on new data mining results or to reflect AIPs that observe common phenomena or areas of interest (i.e. the members of a Theme collection have a common theme). To create a Theme Collection, an archive would create a Collection Descriptor that did not have an associated AIC. The Collection Descriptor could have a customized Associated Member Description that documented the newly mined description data for each member AIP. A specialized finding aid could use this new Associated Member Description in conjunction with existing Associated Descriptions in the Package Descriptor information of each member AIP to locate AIPs of interest to the user. The Package Descriptors of contained AIPs would also supply an automated ordering aid with a method to order the Information Packages of interest to the consumer. Examples of a Theme Collection in a digital movies OAIS might be a new arrivals collection or a twenty most popular titles collection which is updated periodically. If an OAIS decides that a Theme collection is valuable enough to be preserved for the long-term, it can store the required Content and PDI in Archival Storage thus creating a new AIC.

Another important application of Theme Collections is the concept of locating some members of a collection which have been scheduled for ingest at a future time. In this case the Finding Aid would allow the consumer visibility of these future AIPs but the Associated description for the Ordering Aid and/or the Retrieval Aid would contain the information that this product was not currently available.

4.2.3 DATA MANAGEMENT DATA

Currently, Package Descriptors are stored in persistent storage such as database management systems to enable easy, flexible access and update to the contained Associated Descriptions. In addition to the Package Descriptors discussed in the previous sections, all the information needed for the operation of an archive would be stored in databases as persistent data classes. Some examples of this data are accounting data for customer billing and authorization, policy data, request tracking logs, subscription data for repeating requests, and statistical data for generating reports to archive management. These classes are intended as examples rather than an exhaustive list of the data required for archive administration. Figure 4-21 illustrates the various types of "data management data" within the OAIS

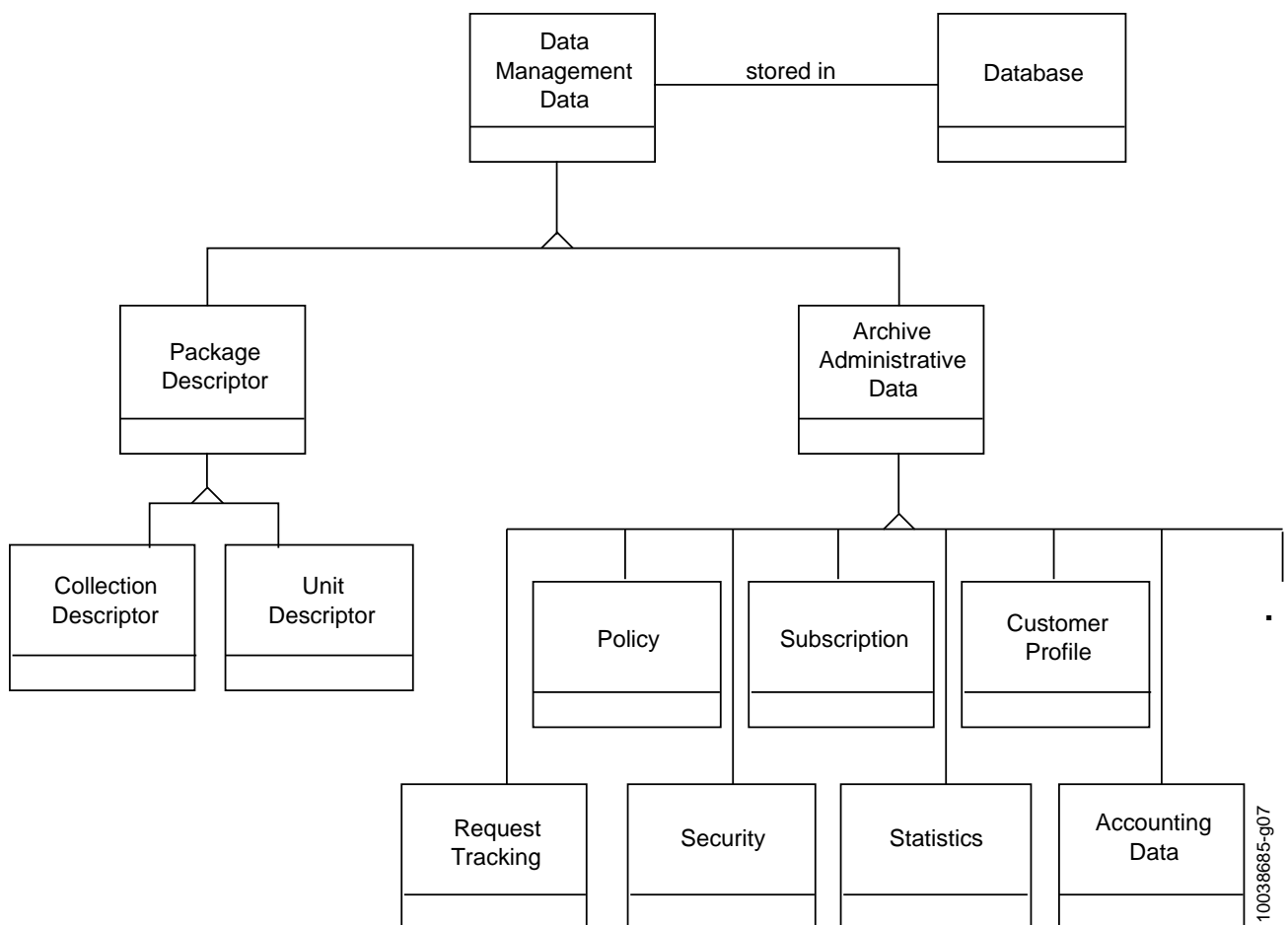


Figure 4-21. Data Management Data

4.3 HIGH LEVEL DATA FLOWS AND TRANSFORMATIONS

Figure 4-19 presents a high level data flow diagram which depicts the principle data flows involved in OAIS operations. These flows do not include administrative flows such as accounting and billing but concentrate on the flows between an OAIS and the other entities in its environment (producers and consumers) and internal OAIS flows that involve Information Packages.

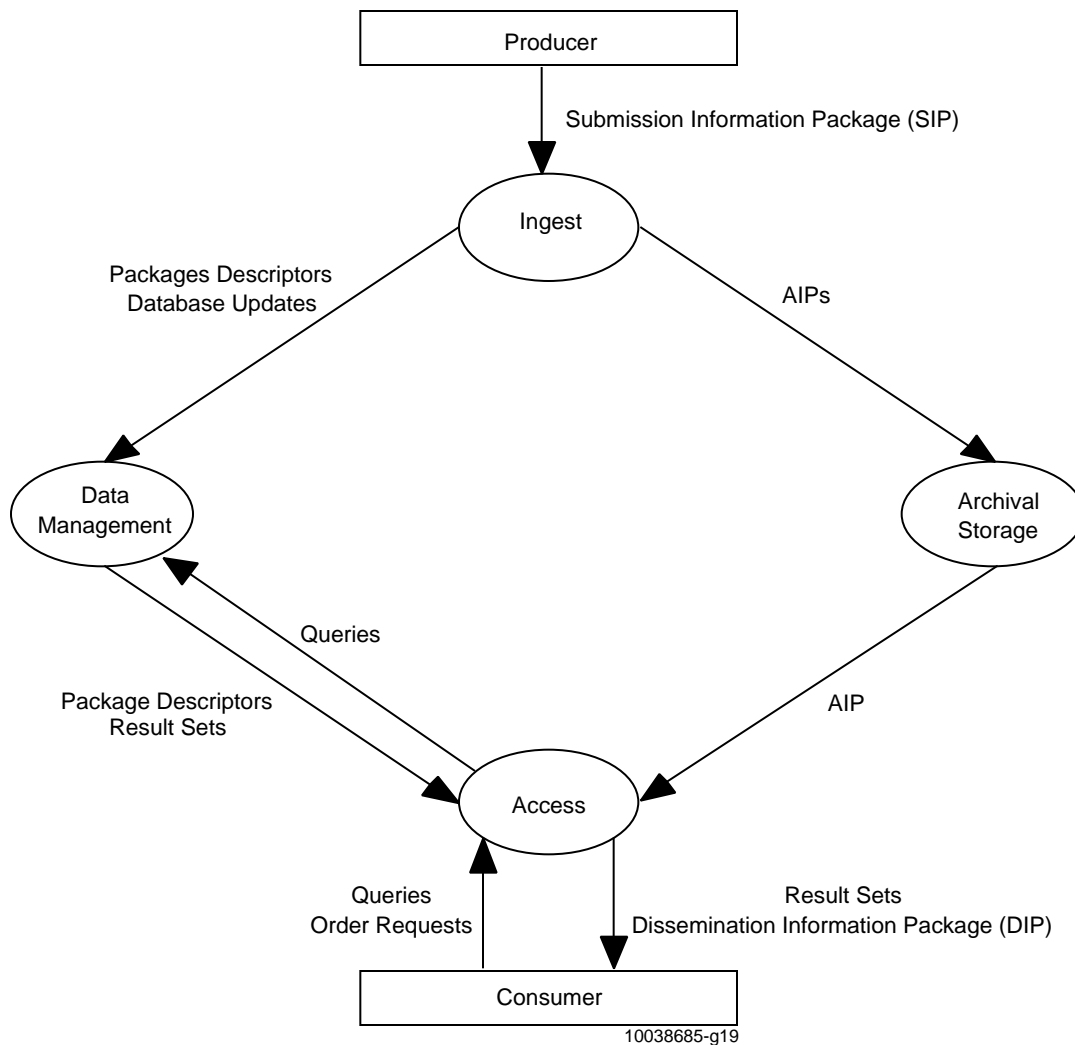


Figure 4-22. High Level Data Flows In an OAIS

When an information object is ingested by an OAIS, various metadata objects are required to assist in the long-term preservation of and access to the information contained in that object. When the information object is stored or migrated, additional metadata objects are created to assist in the preservation and access process. The classes of these metadata objects are

discussed in section 4.2 of this document and form the logical information model of the OAIS. This section discusses the range of potential transformations (both logical and physical) that may occur as an Information Package (IP) passes through the functional areas of the OAIS discussed in section 4.1 of this document.

It is important to note that at each point in these data flows, there is a complete Information Package consisting of Content Objects (e.g., images from a spacecraft instrument) and information describing those content object. There are three separate subtypes of IPs (i.e., the Submission Information Package, the Archive Information Package, and a Distribution Information Package) but the information provided in and with the original submission can be preserved in any of these packages.

4.3.1 DATA TRANSFORMATIONS IN THE PRODUCER ENTITY

The data within the data producer entity are private and may be in any format the producer desired. However, when the decision is made to store the data in an OAIS, the scientific investigators who are responsible for the data meet with archivists to negotiate a Submission Agreement.. This agreement spells out the content and format of the Submission Information Package(SIP). The SIP is an Information Package that is provided to the OAIS by the producer. The SIP consists of the scientific data plus the data that is necessary to assure that those data can be maintained by the OAIS and that the data can be interpreted and used by scientists who withdraw them from the OAIS at some time far in the future. These SIPs are periodically transferred to the OAIS in a Data Delivery Session. The number of data delivery sessions between an OAIS and a data producer can range from a single session in the transfer a final data product to multiple sessions a day in the case of active OAIS which store data for experiments which are still in process. The data delivery session can be logically viewed as sets of content data objects and description objects, although physically the description or metadata can be included in the digital objects (i.e., self describing objects) or divided into many separate descriptive items. In addition to the logical view of data (the SIP), the specification of a data delivery session must also include the mapping of the objects to the media on which they are delivered. This mapping includes the encoding of the object and description and the allocation of logical objects to files.

4.3.2. DATA TRANSFORMATIONS IN THE INGEST FUNCTIONAL AREA

Once the SIP are within the OAIS, their form and content may change. An OAIS is not always required to retain the information submitted to it in precisely the same format as in the SIP. Indeed, preserving the original information exactly as submitted may not be desirable. For example, the computer medium on which our images are recorded may become obsolete, and the images may need to be copied to a more modern medium. In addition, some types of information such as the Reference ID used to locate the Package within the OAIS will not be available to the producer and must be input after ingest to the OAIS.

The mapping between SIPs and AIPs is not one-one.. Here are some examples:

- *One SIP - One AIP* —A government agency is ready to archive its electronic records from the previous fiscal year. All of the year's records are placed onto magnetic tapes that are submitted as one SIP. The archive stores the tapes together as a single AIP.
- *Many SIPs - One AIP* — A satellite sensor makes observations of the Earth over a period of one year. Every week all of the latest sensor data are submitted to the archive as a SIP. The archive has a single AIP containing all of the sensor's observations and the latest SIP is merged into that AIP. The PDI data for the AIP is sent after the last sensor data for the year has been received.
- *One SIP - Many AIPs* — A company submits financial records to an archive as one SIP. The archive chooses to store this information as two AIPs: one which contains public information and the other which contains sensitive information. This makes it easier for the archive to manage access to the information.
- *Many SIPs - Many AIPs* — An oil and gas company collects information on its wells. Every year it submits a SIP containing all of the well status information to an archive. The archive maintains one AIP for each oil or gas field and breaks out the information on each well to the proper AIP based upon its geographic coordinates.
- *One SIP - No AIPs*— An investigator (either within or outside of an archive) creates a new algorithm for detecting hurricanes in images. He runs this algorithm over all the images contained in an archive. This data is combined into either a new Associated Description or a set of Package Descriptor updates which is input as an SIP.

The ingest process transforms the SIPs received in the data delivery session into a set of AIPs and Package Descriptors which can be stored and accepted by the Storage and Data Management functional entities. The complexity of this ingest process can vary greatly from OAIS to OAIS or from producer to producer within an OAIS. The simplest form of the process involves removing the Content Information, PDI and Package Descriptors from the producer transfer media and queuing them for storage by the storage and data management functions. In more complex cases, the PDI and Package Descriptors may have to be extracted from the Content Information or input by OAIS personnel during the ingest function; the encoding of the information objects or their allocation to files may have to be changed; in the most extreme case, the granularity of the Content Information may be changed, and the OAIS must generate new PDI and Package Descriptors reflecting the newly generated information objects. When many SIPs are required for the creation of one AIP, the Ingest functional area will provide staging storage for the SIPs until all the SIPs required for the AIP arrive.

In addition, the Ingest functional entity will classify incoming information objects and determine in what existing collection or collections each object belongs and will create messages to update the appropriate collections Descriptors after the AIPs are stored.

OAISs and external organizations may provide additional Associated Descriptions and finding aids that allow alternative access paths to the information objects of interest. As data mining technologies become more mature, it is likely that researchers will develop new and fundamentally different access patterns to information objects. It is important that an OAIS's Ingest and internal data models are sufficiently flexible to incorporate these new descriptions so the general user community can benefit from the research efforts. A good example of this type of new associated description are a phenomenology database in Earth Observation,

which allows users to obtain data for a desired event such as a hurricane, or volcano eruption from many instruments with a single query. It is important to note that such finding aids may become obsolete unless the data they require are preserved as parts of the AIPs they access.

During the ingest process, the OAIS must be highly aware not to unintentionally modify any of the information content in the Producer view. The OAIS is advised to save the producer media or copy the media into long-term storage as an ultimate reference if needed. It should be recognized that the saving of the producer media will not be permanent because of the issues discussed in Sections 2 and 4.2.2 of this document.

4.3.3. DATA TRANSFORMATIONS BY THE ARCHIVAL STORAGE AND DATA MANAGEMENT FUNCTIONAL AREAS

The Archival Storage functional entity take the AIPs produced by the ingest process and merge it into the permanent archive holdings. The Data Management functional entity takes the Package Descriptors produced by Ingest and augment the existing collection of Descriptors to include their contents. The logical model of the ingested data should already be mapped into the logical model of the archives holdings, so the major transformation that occurs in this step is mapping the acquisition session from the ingest physical data model, which will tend to be on staging storage, to the permanent storage of the OAIS, which could range from database management systems (DBMSs) to hierarchical file management systems (HFMS), or any mixture of the above. It is expected that the Ingest functional entity will coordinate the updates between Data Management and Ingest Storage and provide appropriate error recovery.

The internal view of the OAIS is the permanent representation of the archived data, so all encoding and mappings must be well documented and understood. The process of transferring the ingest objects is frequently by a software process such as an HFMS driver or a DBMS. In this case, it is the responsibility of the OAIS to maintain an active copy of the software or careful documentation of the internal formats so the data can be transferred to other systems in the future without loss of information.

4.3.4. DATA FLOWS AND TRANSFORMATIONS IN THE ACCESS/DISSEMINATION FUNCTIONAL AREA

When a Consumer wishes to use the data within the OAIS, he may use a finding aid to locate information of interest. These finding aids present data consumers with the logical view of the OAIS so the consumers can decide which information objects they wish to acquire. At a minimum, the access view is the high-level logical view of collections described in section 4.2.3. In most cases, the OAIS will have spent significant time and effort developing associated description and finding aids such as catalogs that will aid the user in locating information objects or collections of interest. The consumer will establish a Search Session with the Access/Dissemination entity. During this Search Session, a Consumer will use the OAIS finding aids to identify and investigate potential holdings of interest. This searching process tends to be iterative, with a user first identifying broad criteria and then refining the criteria on the basis of previous search results. When the user has identified candidate objects

of interest he may use more sophisticated visualization aids such as browse image viewers or animation to further refine his result set.

Once the Consumer identifies the OAIS holdings he wishes to acquire, he must issue an order request to the OAIS to acquire the data. The data consumer produces a logical view of the desired information objects and Descriptors to be included in the Dissemination Information Package. At this point, the consumer issues an order request for this DIP that triggers the negotiation of a request agreement with the customer in which the physical details of the Data Dissemination Session such as media type and object format are specified. This process may involve no visible interaction between the consumer and the archive if adequate defaults exist. This order can also specify any transformations the Consumer wishes applied to the AIPs in creating the Dissemination Information Package (DIP).

The Access/Dissemination functional area then records the Order Request in the Data Management functional area. When the conditions required to request a dissemination based on an recorded Order are met (note for many Orders these conditions are met immediately) the Data Management functional area sends a dissemination request to Access/Dissemination. Access/Dissemination then contacts the Storage and Data Management functional areas and requests the AIPs and associated Package Descriptors necessary to populate the DIP requested by the consumer. The Storage and Data Management functional areas create copies of the requested objects in staging storage.

Access/Dissemination then transforms this set of the AIPs and associated Package Descriptors into a set of DIPs and stores those DIPs onto physical distribution (either physical or communications) media to be delivered to the data consumer in a Data Dissemination Session. The complexity of this transformation process can differ greatly on the basis of the level of processing services offered by the OAIS and requested by the data consumer in his order. In the simplest case, the DIP contains duplicates of the AIPs and associated Package Descriptors of interest from storage and data management function. In more complex cases, the desired Content Information may have to be extracted from the information objects or inserted into self-describing information objects, and the encoding of the information objects or their allocation to physical files may have to be changed. In the most extreme case, when the OAIS supports subsetting services, the granularity of the of the information objects may be changed, and the Dissemination process may generate AIPs and associated Package Descriptors reflecting the newly granularity. In most cases the mapping between DIPs and AIPs is 1:1 however the use of subsetting services and other product processing options could create many DIPs from 1 AIP.

5 DIGITAL MIGRATION PERSPECTIVES

The fast-changing nature of the computer industry and the ephemeral nature of electronic data storage media are at odds with the key purpose of an archive: to preserve its digital information over a long period of time. No matter how well an archive maintains its current holdings, it is likely that over time much of the stored information will need to migrate to different media or to a different hardware or software environment to remain accessible. Today's digital data storage media can typically be kept at most a few decades before the probability of irreversible loss of data becomes too high to ignore. Further, the rapid pace of technology evolution makes many systems much less cost-effective after only a few years.

Digital Migration is defined to be the transfer of digital information, while intending to preserve it, within the OAIS. It is distinguished from transfers in general by its focus on the preservation of the full information content, and by its internal OAIS perspective. However the concepts involved have application to other transfers as well.

This Section uses the AIP concepts of Section 4.2 and explores their relationships to OAIS preservation approaches in the face of decaying media, changing technologies, and changing service requirements. It focuses on the preservation of the AIP's Content Information and Preservation Description Information, and the role of the Packaging Information in supporting this preservation within the OAIS. First the motivators for Digital Migration are identified. Then, the AIP concepts are expanded to identify key features which influence, and are impacted by, migrations. A model of the media, hardware, and software environment is provided to serve as a basis for examining the ways in which the key features of the AIPs may be allocated, and three major categories of allocation are identified and described. With this background, two basic types of Digital Migrations are identified and explained using examples from the three major categories of allocation. Finally, the issues if AIP versions is addressed and related to AIP concepts and migration approaches.

5.1 DIGITAL MIGRATION MOTIVATORS

Three major motivators are seen to drive Digital Migrations of AIPs within an OAIS. These are:

- **Media Decay:** Digital media, over time, become increasingly unreliable as secure preservers of bits. Even those that are used with some level of error correction eventually need to be replaced. The net result of media decay is that AIP information must be moved to newer media.
- **Incorporation of New Technologies:** The rapid pace of hardware and software evolution provides an OAIS with greatly increasing storage capacities and transfer bandwidths at reducing costs. It also drives the obsolescence of some media types well before they have time to decay. In addition, improved AIP packaging designs which are less dependent on underlying media and supporting systems may be recognized. To remain cost-effective, an OAIS must take advantage of these technologies. Depending on the particular technologies involved, the AIP information may have to be moved to new media types not previously supported and it may have to revise its AIP implementations

to take advantage of the new technologies.

- **New Consumer-Service Requirements:** The Consumers of an OAIS also experience the benefits of new technologies and consequently raise their expectations of the types and levels of service they expect from an OAIS. These increased services may require new forms of DIPs, which in turn may drive an OAIS to hold new forms of AIPs to reduce output conversions. Additionally, AIPs typically go through popularity swings and the OAIS may need to provide different levels of access performance to meet Consumer demands over time. This is likely to be satisfied by moving some AIPs to different media that provide increased or decreased levels of access performance. Finally, the designated community for a given AIP may be broadened, resulting in the need to revise AIP forms so as to be understandable and usable by this broader community. All of these can result in the migration of AIPs within an OAIS.

Digital Migrations are time consuming, costly, and expose the OAIS to greatly increased probabilities of information loss. Therefore an OAIS has a strong incentive to avoid migrations when possible, and otherwise to reduce manual involvement to save cost. The ways in which AIPs are implemented will have a major influence on both the level of automation and the probability of information loss. Good AIP designs can both increase migration automation and reduce information loss probabilities. To better understand the impacts of these factors on AIP migrations it is necessary to first extend the AIP model to slightly greater detail.

5.2 AIP OBJECTS : COMPONENTS, POINTERS, AND PACKAGE STRUCTURES

The AIP information model has been presented in Section 4.2. Conceptually the AIP Packaging Information identifies and relates the various objects making up the Content Information and the PDI. It is necessary to model these Information Objects in greater detail to better discern the migration issues.

The Content Information and PDI Information Objects are further modeled in terms of **Components** and **Pointers**. The Packaging Information is further modeled in terms of **Package Structures** and **Pointers**. It will be shown that Pointers play key roles in constructing AIPs and in defining dependencies between the Packaging Information and the contained Content Information and PDI. Depending on their characteristics, they may greatly simplify or complicate a given migration. The next section examines the modeling of Content Information and PDI, and in the process drives out more detailed requirements for the Packaging Information. The subsequent section (5.2.2) takes these requirements and models the Packaging Information in detail, showing how it supports the Content Information and PDI.

5.2.1 CONTENT INFORMATION AND PDI COMPONENTS AND POINTERS

Components are defined to be independent, separately identifiable, parts of the Content Information or PDI Information Objects. They are locatable because they have an implicit or explicit identification and can be 'pointed to' by one or more Pointers. A Component may be implemented in many ways, such as the contiguous sequence of bits in the content of a file,

or as a single hardcopy document. The file content may be located by virtue of a file and directory pathname, or by a file count on some sequential medium. The hardcopy document may be located by virtue of its storage in an identified storage bin. A digital Component may span multiple files or many may exist within a single file, as long as each has an identification and can be pointed to by a Pointer. A Component may also contain embedded Pointers which therefore allows it to describe relationships among other Components.

A Pointer identifies one or more Components. A Pointer may appear as a pathname which has multiple structural parts used to identify different parts of the path, or it may appear as a simple name which is associated with mechanisms for resolving it into a corresponding pathname. A pointer to the file in the example above may appear as a pathname which includes physical site, computational node, device, directory path, and file name. Various mechanisms, including the target medium's directory and file structure, support the location function which is able to arrive at the identified file content using the Pointer.

The Content Information and PDI objects will always be modeled as multiple Components because at the very minimum there must be a distinction between the Content Information and the PDI. Digital Content Information will itself appear as multiple Components where at least one is needed for the primary Digital Object and others are needed for the Representation Information object(s). There are two basic relationships between Components and Pointers from which all Content Information and PDI can be constructed. These are:

- Components with Representation Pointers
- Components with Embedded Pointers

5.2.1.1 Components with Representation Pointers

Many Components, and all digital Components, need one or more Representation Information Components to provide additional meaning. This relationship is satisfied by associating a Representation Pointer with a Component, where the Pointer 'points to' one or more Representation Information Components. An example is shown in Figure 5-1 where a primary Digital Object (Component 1) of some Content Information needs Representation Information Component 2 to understand its bits. However Component 2 is also digital and it needs Representation Information Component 3 to understand its bits. Component 3 is a hardcopy document stored in an identified storage bin.

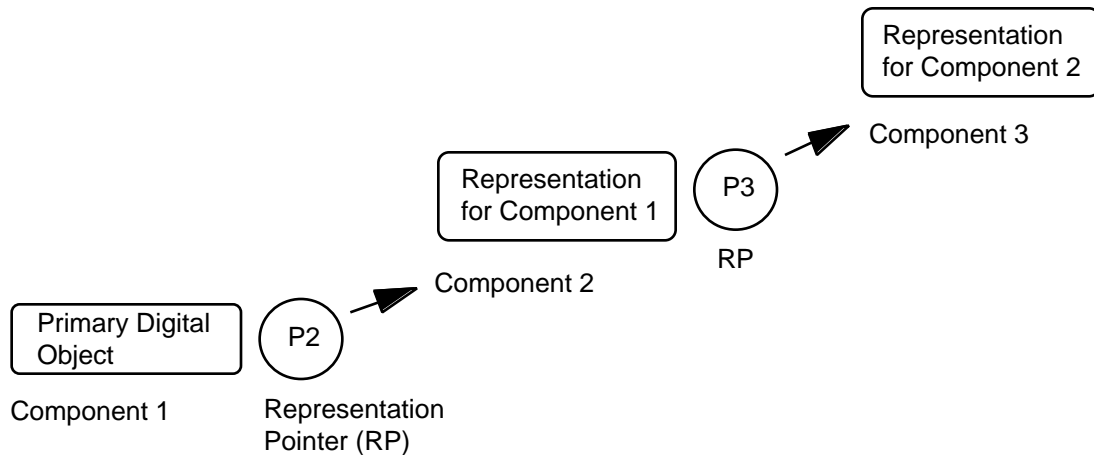


Figure 5-1: Components related using Representation Pointers

For this example, the Content Information consists of Components 1, 2, and 3 together with Representation Pointers P2 and P3. The functionality required of the Packaging Information for these types of relationships can be inferred to be the following:

- Encapsulates and identifies the Representation Pointers and the Components, and it relates the Pointers to their Components
- Encapsulates and identifies, as a single collection, the Components and Pointers comprising the Content Information.

5.2.1.2 Components with Embedded Pointers

Content Information can be comprised of Components whose relationships are infinitely variable. Therefore there is the general requirement for one Component to be able to describe the relationships among any number of other components. This requires that the describing Component contain one or more Embedded Pointers which refer to other Components. An example is shown in Figure 5-2 where Component 1 specifies relationships among Component 2 and Component 3.

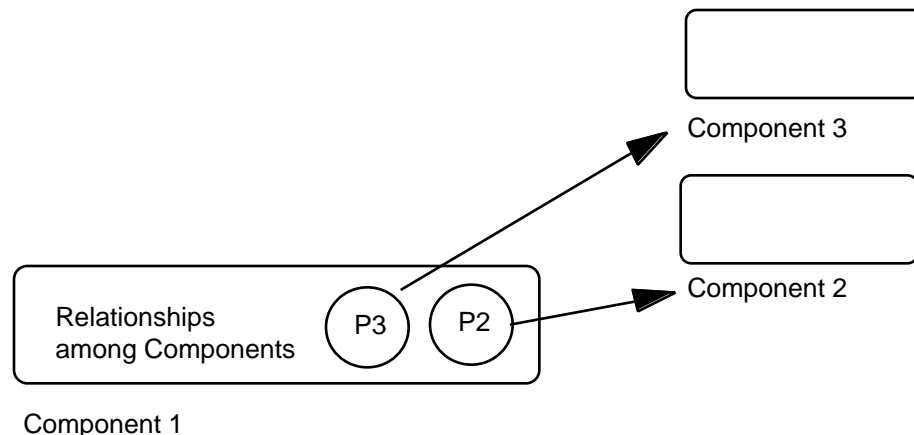


Figure 5-2: Component related using Embedded Pointers

Component 1 may be a text document that describes how Component 2 is used as calibration information to convert Component 3 into physically meaningful numbers. If any of these are digital components, they will need to have relationships with Representation Pointers and additional Components as described above.

For this example, the Content Information consists of Components 1, 2, and 3. Component 1 already includes the Embedded Pointers P2 and P3. The functionality required of the Packaging Information for these types of relationships can be inferred to be the following which is a subset of that given above:

- Encapsulates and identifies the Components
- Encapsulates and identifies, as a single collection, the Components comprising the Content Information.

5.2.1.3 Packaging Requirements

The two sections above addressed Components and Pointers in the context of Content Information. Their application will be the same for the PDI objects. However the Packaging Information will have to also separately identify these objects and also identify the entire collection of objects. Therefore the Packaging Information must provide the following functionality:

- Encapsulates and identifies the Representation Pointers and the Components, and it relates the Pointers to their Components
- Encapsulates and identifies, as single collections, the Components and Pointers comprising the Content Information object and the PDI objects.
- Encapsulates and identifies, as a single collection, all the objects.

Also, although not called out above, the Packaging Information must identify, for each object, which Component should be the starting point for understanding the Information Object.

5.2.2 PACKAGING INFORMATION STRUCTURES AND POINTERS

The requirements for the Packaging Information have been delineated immediately above. They are met through the use of modeling concepts called Package Structures and optional Package Pointers. The Package Structure encapsulates and identifies Components and Representation Pointers. This is shown schematically in Figure 5-3.

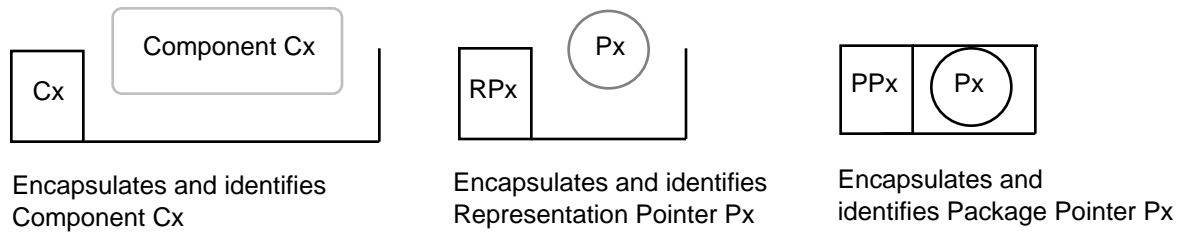


Figure 5-3: Schematics for identifying Package Structure types

An example of a Package Structure implementation that can encapsulate and identify a Component or a Representation Pointer is a file structure with an associated directory structure. The file structure encapsulates the file content, which could be a Component (e.g., Cx) or a Representation Pointer (e.g., Px). The directory structure and file name support an identification of the file content, which is represented by Cx or RPx in Figure 5-3.

The Package Structure also needs to be able to encapsulate and identify multiple Components which may already be in a Package Structure. When the Package Structure is not contiguous, it needs to be able to point to additional Package Structure. This type of Pointer is called a Package Pointer and is also shown schematically in Figure 5-3. Following the example, an implementation of a Package Pointer could be another file structure whose name supports the identification PPx and whose content holds the Package Pointer Px.

Note that the actual value of a Pointer may be the same regardless of the role that it plays, whether Representation Pointer, Embedded Pointer, or Package Pointer. The difference is where it is found in relation to Components and Package Structures.

As an example of more complex Packaging Information, Figure 5-4 provides a schematic view of packaging the three Components presented in Figure 5-1.

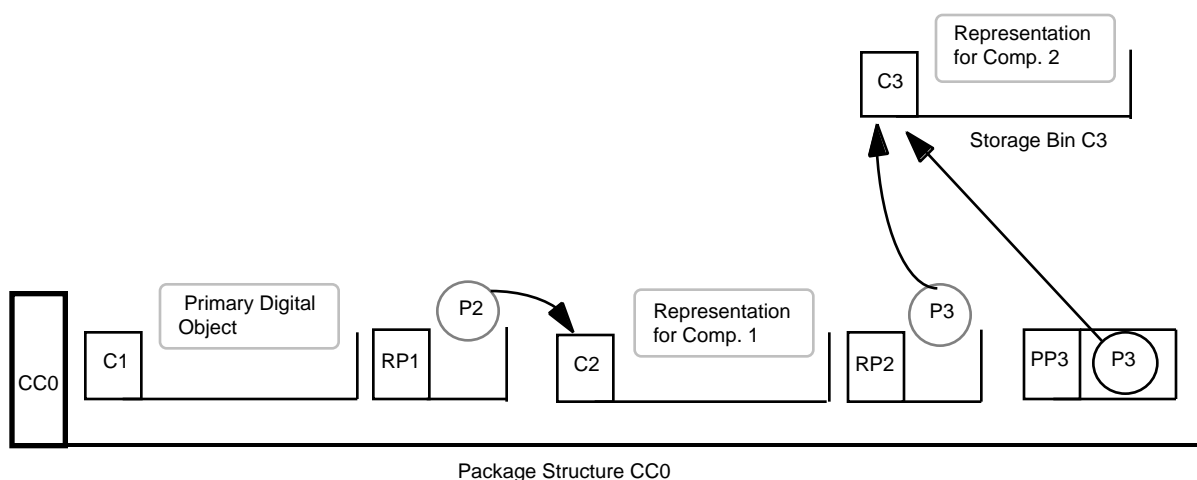


Figure 5-4: Packaging Information for example Content Information of Figure 5-1

In this example Package Structure C1 holds the primary Digital Object (Component 1), Package Structure RP1 holds the Representation Pointer P2 which points to C2 (Component 2). Its Representation Pointer P3 is held by Package Structure RP2. Package Structure PP3 hold the Package Pointer P3. All of these Package Structures are encapsulated and identified by Package Structure CC0. An implementation of CC0 might be a directory structure which contains file structures that hold the four Content Information Components “Primary Digital Object”, P2, “Representation for Comp. 1”, and P3. A fifth file structure might implement PP3 and also hold P3. Note that P3, as held by PP3, is performing the function of logically incorporating the Storage Bin C3 with the Packaging Structure CC0. However the Representation Pointer P3 is performing the function of identifying the Component in the storage bin as the one needed to understand Component 2 (C2).

Note that all the Pointers actually point to Package Structure and not to Components directly. Pointers from outside CC0 that wish to point to Component C2 will have to point to CC0 on the way to C2. Therefore Representation Pointers and Embedded Pointers actually straddle the boundary between “Content Information and PDI” and “Packaging Information”. Since many Digital Migrations will be motivated by other than requirements to reformat the Content Information, their primary objective will be to ensure there is no change to the “Content Information and PDI”. However their Representation Pointers and Embedded Pointers may need alteration as the Package Structure changes. The nature of this interaction depends on the implementation characteristics of both the Pointers and the Package Structure. The next Section models implementation characteristics of Pointers.

5.2.3 MODELING POINTERS

Pointers may be classified as being either a **Local Pointer** or a **Global Pointer**. A Local Pointer is one that is recognized only within the OAIS and therefore points to Components held within the OAIS. A Global Pointer is recognized both within and external to an OAIS. Global Pointers are further classified as being an **Internal Pointer** or an **External Pointer**. An Internal Pointer is a Global Pointer that points to Components held within the OAIS. An External Pointer is a Global Pointer that points to Components held outside the OAIS.

Any Pointer that is implemented as a pathname will need updating if any of the structural parts of the pathname no longer give the correct path. Therefore it is highly desirable to implement Pointers that are not in a pathname form, but rather serve as Universal Identifiers (constant identifiers) of the Component to which they point. This moves the problem from one of updating names buried in Components and of altering Content Information that is protected by Fixity information to the problem of exercising the underlying mechanisms by which Universal Identifiers retain their integrity. This is a much more easily managed scenario because such updates are part of the design of Universal Identifiers. The net result is much less risk of loss of information by the OAIS.

To the extent that Global Pointers are used, instead of Local Pointers, the transfer of AIPs to external Customers, in the form of DIPs, may be eased because not all of the Components need to be transferred. Global Pointers still present in the DIP, that point back to OAIS holdings, will still be valid. This allows the Consumer to selectively obtain Components as

needed. However, if the Global Pointers are also Universal Identifiers, there is a further reduction in the effort needed to maintain Pointer integrity. While Global Pointers are not strictly a Digital Migration issue within the OAIS, this issue does impact the implementation strategies of Pointers.

An OAIS may wish to cooperate with another OAIS in the maintenance of common Components. Through the use of External (Global) pointers an OAIS may point to Components under the control of another OAIS. Such Components are not affected by Digital Migrations within the first OAIS. However, there needs to be a strong agreement that any such shared Components will not be altered unilaterally.

To further explore the impacts of implementation choices for the Package Structure, it is first necessary to develop a simple model of the environment in which the Package Structure may exist.

5.3 DIGITAL MIGRATION ENVIRONMENT MODEL

This Section provides a simple model of the hardware and software environment in which Digital Migration is likely to take place. It addresses physical media up through a typical File Storage Management System (FSMS). It does not attempt to address the additional complexity that a DBMS would add because this is not necessary to make the key points.

Digital information resides on physical media within the environment. It is assumed to be written to the media by a writing device controlled by driver software, and it is read by driver software controlling a reading device. The reading and writing devices may, or may not, be the same device.

File management software, working with the driver software, writes and/or reads a particular **Media Format**. This format is independent of any application data that will be stored on the media. Examples of such Media Formats include end-of-files written on unlabeled 9-track tape, ANSI standard labels written on 9-track tape, ISO-9660 volume/directory structures on CD-ROMs, mass storage vendor-specific formats on Digital Linear Tapes (DLTs), and VMS file formats on tape. As a short hand, all these formats are considered to be written by a corresponding FSMS. Each FSMS provides a set of services to applications that want to store information on the media. The information to be stored, which will be carried by the Media Format, is viewed as being given by an application to the FSMS in the form of a set of attributes. The primary attribute is taken to be a sequence of bytes which will be called the **File Content**. For each file, and for collections of files including the media volume itself, there are **Supporting Attributes**.

At a minimum, all the FSMSs are assumed to provide the ability to store and retrieve multiple sets of application-provided byte sequences, which are stored as File Content. Most of the FSMS also provide a range of related services which accept additional application provided information and return it upon request. These are the Supporting Attributes. Examples of Supporting Attributes include volume names, directory names, file names, file content attributes, creation dates, record delimiters, and access permissions. These additional

services, and their Supporting Attributes, are not standardized across media types, and often vary across vendor FSMSs dealing with the same media type. This poses a challenge for information migration across these FSMSs and the underlying media. A schematic of an example file, composed of Supporting Attributes and File Content, is shown in Figure 5-2.



Where file content of file x is:

- concatenation of byte sequences A, B, and C

Where supporting attributes of file x are:

- a is file name
- b is creation date
- c is read/write permission
- r is record delimiter

Figure 5-2: Schematic of an example file

The Media Format can be viewed roughly as a layer of files spread, in general, across media volumes as shown schematically in Figure 5-3.

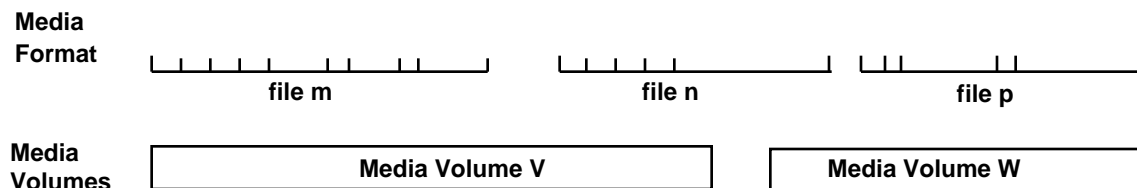


Figure 5-3: Media Format on media volumes

In general, the Media Formats depend not only on the types of media volumes involved, but also on the FSMS and operating systems used to write the Media Formats. Therefore the ways in which the File Content and Supporting Attributes of FSMSs are used to support AIP Components, Pointers, and Package Structures can play a significant role in the ease or difficulty of AIP Digital Migration.

5.4 CATEGORIES OF PRESERVATION

Any given instance of an AIP will have its Components, Pointers, and Package Structures mapped to the Media Format on one or more Media Volumes. From the perspective of preserving digital Content Information, there are three primary mapping categories that can be identified and characterized as:

- Media Format Preservation
- Media Supporting Attribute Preservation
- Media File Content Preservation

5.4.1 MEDIA FORMAT PRESERVATION

This characterizes the situation where a specific media format, including all its bits, is declared to be the Content Information's primary Digital Object. This might occur when a particular piece of media participated in an important historical event, or when there was a sizable community that expected not to have to alter any of its access software. This means that the underlying media volume is a part of the Package Structure. Pointers to this volume depend on its location and access device. The remainder of the Components, Package Structure, and Pointers will reside on other media.

Generally, Media Format Preservation means that the replacement media will have to be of the same type as the original because of the often strong coupling of media format to media volume type. For example, the declaration of Content Information as including the ISO-9660 media format on a CD-ROM means that this Component can readily be migrated to another CD-ROM, but could not easily be migrated to Digital Linear Tape (DLT) while maintaining cost-effective access because this media type does not support ISO-9660 directly. Special tools would be needed to extract all the ISO-9660 bits and most likely they would be stored as a single large file on the DLT.

5.4.2 MEDIA SUPPORTING-ATTRIBUTE PRESERVATION

This characterizes the situation where some of the Supporting Attributes, together with some of the File Content of a Media Format, are used to hold the Components and related Pointers of the Content Information. The PDI may be found in other File Contents and Supporting Attributes on the same or other media. The Package Structure will consist of the Media Format bits that encapsulate and identify the File Contents and Supporting Attributes of the Digital Object, the information that specifies which File Contents and Supporting Attributes are used for the Digital Object, and the Representation Information for the Media Format. Also included in the Package Structure will be the underlying physical medium which is holding the Packaging Information bits.

The use of Supporting Attributes, in addition to File Content, for the Components and related Pointers makes the migration of this information to new media types and their new Media Formats more complex because there may not be a simple mapping to the new Supporting Attributes that also preserves all information, including the relationships.

5.4.3 MEDIA FILE CONTENT PRESERVATION

This characterizes the situation where only the File Contents of a Media Format are used to hold the Components and related Pointers.. In other words, these bits are not found in any related file attributes, including the file name. The Package Structure is the Media Format

bits, together with its Representation Information, that encapsulate and identify the File Contents. Also included in the Packaging Information will be the underlying physical media, which is a part of the Package Structure, that is holding the Packaging Information bits.

The use of only File Content for holding Components and related Pointers makes Digital Migration less complex because most Media Formats will support this encapsulation. However there are still concerns associated with embedded pathname Pointers. Such Pointers involve the Package Structure identification within a Component and place a dependence of the Content Information or PDI on the local Packaging Information. This type of coupling between the Content Information and the Packaging Information may require alteration of the Content Information during a migration.

5.5 MIGRATION TYPES

Based on the models and concepts above, it is possible to identify two primary digital migration types. The primary types are:

- **Repackaging:** A Digital Migration where there is some change in the Packaging Information.
- **Transmutation:** A Digital Migration where there is some change in the Content Information or PDI.

5.5.1 REPACKAGING

Virtually all Digital Migrations involve some changes in the Packaging Information, and many simultaneously involve some changes in the Content Information or Preservation Description Information. In other words, many Repackaging migrations also involve Transmutation. This is referred to as Repackaging with Transmutation.. A special type of Repackaging occurs when the ONLY change to the Packaging Information is to replace a physical media instance with another of the same type. This is referred to as **Replication**. In this situation, there is also no change to any of the software needed to read the media format, or any of its attributes. This is very closely related to the concept of **Media Refreshment**, but there can be a difference. With Media Refreshment, the change refers to replacing a physical media instance with another of the same type. Depending on how this is done, the result may or may not show any bit differences in comparison with the original media instance. Even though the same software may read and use the new media and its media format, there may be some slight differences in some of the bits (e.g., some bits set by the media manufacture). Therefore a Media Refreshment will be a form of Repackaging, but if there are any bit differences it is not Replication. This distinction can be important, such as when the Media Format has been declared to be Content Information. To assist in clarifying Repackaging, examples are given for each of the three categories of preservation given in Section 5.4.

5.5.2 REPACKAGING EXAMPLES

5.5.2.1 Media Format Preservation

In this example, the Media Format of a CD-ROM, which therefore includes all the bits of the CD-ROM, has been declared to be part of the Content Information. The Packaging Information includes the physical CD-ROM volume.

Transferring all the bits to a new CD-ROM is an example of both Repackaging and Replication. However, using common utilities to make a copy to the new CD-ROM volume may not preserve all the bits. In this case, the Media Format will not have been preserved and therefore the Content Information will have undergone a Transmutation.

5.5.2.2 Media Supporting-Attribute Preservation

In this example, the Content Information is declared to include be the content of two CD-ROM files together with the file names and the directory name of the directory that contains the two files. The Packaging Information includes the physical CD-ROM volume and the media format bits (and Representation Information) that establishes the directory structure and file structure, including the encapsulation of the directory name and file name Supporting Attributes.

Transferring the directory and files to a new CD-ROM of the same type where there is no change in any of the bits on the medium is an example of Replication. If the overall directory structure on the new medium is different, but the representations of the directory name and two file names associated with the two files has not changed, then a Repackaging (but not Replication) has taken place because there is no change in Content Information. However, if there has been an alteration in the representation of the directory or file names, then Replication with Transmutation has taken place.

5.5.2.3 Media File-Content Preservation

In this example, the Content Information is declared to include the File Content of three files on a CD-ROM. The Packaging Information includes the physical CD-ROM volume and the media format bits (and Representation Information) that establishes both file structure and the identification of each File Content. The directory structure and the file names support Pointers to the File Content, and are part of the Packaging Information.

Transferring the three files to a new CD-ROM of the same type in such a way that there is no change in any of the bits on the medium is an example of Replication. If the overall directory structure on the new medium is different, even to the extent of using different file names, then a Repackaging (but not Replication) has taken place because there is no change in Content Information. However, if one of the transferred files included an embedded Local Pointer in the form of a pathname to the other file and the directory structure has changed, then the Content Information needs to be updated with the new pathname to the other file. This is a case of Repackaging with Transmutation. On the other hand, if the embedded Pointer was a Universal Identifier, then the association to the new location of the file is handled through the mechanisms which support the Universal Identifier and there is not change to the Content Information. This is a case of Repackaging.

5.5.3 TRANSMUTATION

Digital Migrations that require some changes to the Content Information or PDI are referred to as Transmutations. These changes may be to some of the bits in the Primary Digital Object of the Content Information or PDI with corresponding changes in the associated Representation Information, or they may only be to some of the Representation Information. In all cases the intent is to provide maximum information preservation because the resulting AIP is intended to be a full replacement for the AIP that is undergoing Transmutation. The new AIP qualifies as a new **Version** of the original AIP.

The Representation Information plays a key role in Transmutations and the impacts of the changes on the Representation Information may be used to categorize the Transmutations. A Representation Information object can be modeled as consisting of a base set of entities, a set of resulting entities, and mapping rules that define the resulting entities and their relationships in terms of the base entities. Using this model of a Representation Information object, two types of Transmutation can be defined: **Reversible Transmutation** and **Non-Reversible Transmutation**. Reversible Transmutation occurs when the new representation defines a set (or a subset) of resulting entities that are semantically-equivalent to the resulting entities defined by the original representation. This means that there is a one-to-one mapping back to the original representation and its set of base entities. An example is replacing a representation that uses the ASCII codes “A through Z” with a representation that uses the UNICODE codes for “A through Z”. The Transmutation will result in the replacement of 7-bit codes with 16-bit codes in the AIP object undergoing change. The reverse Transmutation can subsequently be performed by replacing the UNICODE codes for “A through Z” with the ASCII codes for “A through Z” and the original AIP is recovered.

Non-Reversible Transmutation occurs when Reversible Transmutation can not be guaranteed. For example, replacing an IBM 7094 floating point value with an IEEE floating point value is a Non-Reversible Transmutation because the resulting entities of these two representations are not semantically equivalent. One will have more precision than the other. However they may be sufficiently equivalent, depending on what the values they represent are being used for, to be effectively interchangeable. If this is the case, Non-Reversible Transmutation effectively preserves the information content. For complex formats, where the meanings and relationships among groups are significant, it may be difficult to establish that Non-Reversible Transmutation has adequately preserved the Content Information. The impacts of the migration drivers are as follows: To assist in clarifying Transmutations, examples are given for each of the three categories of preservation given in Section 5.4.

5.5.4 TRANSMUTATION EXAMPLES

5.5.4.1 Media Format Preservation

In this example, the Media Format of a CD-ROM, which therefore includes all the bits of the CD-ROM, has been declared to be part of the Content Information. The Packaging

Information includes the physical CD-ROM volume.

Over time the requirement to preserve all the bits associated with the Media Format is relaxed to a requirement to preserve the Media Format functionality and the bits in the File Content. A new media type with a new Media Format, which supports at least the same Supporting Attributes, File Contents, and their inter-relationships as the original, is used to hold the values of the original CD-ROMs Supporting Attributes and File Contents. The result is a Transmutation that, in all probability, is Non-Reversible because there is unlikely to be a set of mapping rules from the new media format type back to the original media format type. The only likely way to get back is to retain all the original media format instances. Keeping the original version of the AIP is a conservative information-preserving approach for the OAIS but has the problem of increasing storage volume and requiring the maintenance of hardware and software to read and write the old media..

5.5.4.2 Media Supporting-Attribute Preservation

In this example, the Content Information is declared to include be the content of two CD-ROM files together with the file names and the directory name of the directory that contains the two files. The Packaging Information includes the physical CD-ROM volume and the media format bits (and Representation Information) that establishes the directory structure and file structure, including the encapsulation of the directory name and file name Supporting Attributes.

The directory and two files are transferred to a new media type whose Media Format supports a directory structure and will hold the File Content bits of the original CD-ROM. However, the new directory structure uses UNICODE to implement the directory names and file names. The result is therefore a Reversible Transmutation because the directory and file names have new representations from which the originals can be derived using a set of mapping rules.

5.5.4.3 Media File-Content Preservation

In this example, the Content Information is declared to include the File Content of three files on a CD-ROM. The Packaging Information includes the physical CD-ROM volume and the media format bits (and Representation Information) that establishes both file structure and the identification of each File Content. The directory structure and the file names support Pointers to the File Content, and are part of the Packaging Information.

In one scenario, the three File Contents are transferred to a DLT, and in the process they are compressed. This transfer is a Transmutation because the compression process has altered the Content Information, and it is a Reversible Transmutation because there is a decompression algorithm that will return the original File Content bits. The relevant Representation Information Components of the original Content Information need to be updated to include the decompression algorithm, and the PDI information needs to be updated, in forming this new AIP version.

In a variation on the above scenario, the format of the third file is also altered before the compression is performed because there are no longer readily available tools to make effective use of the third file's File Content in its current form. The new format, which is in common use, employs a different data model from that of the original format and there are many ways in which the information may be mapped into the new format. This mapping must be carefully done to ensure there is no significant information loss to the designated community. This mapping must be included in the PDI, and of course the Representation Information describing the new format will replace that which was describing the previous format, for the new AIP version.

5.6 DISTINGUISHING AIP VERSIONS, EDITIONS AND DERIVED AIPS

A Digital Migration that is only Repackaging is not considered to create a new AIP version and it is not required that its PDI needs to be updated. In other words, the AIP version is considered to be independent of Repackaging that does not affect the Content Information or PDI. This does not mean that the OAIS does not track such Repackaging, only that it is not required to update the PDI as part of such tracking.

A Digital Migration that involves Transmutation results in a new version of the AIP as defined in Section 5.5. In this case, the PDI needs to be updated to identify the source AIP and its version, and just what was done and why. The new AIP is viewed as a replacement for the source AIP where the information has been preserved to the maximum extent practical.

An AIP may, in some environments, be subject to upgrading or improvement over time. This is not a Digital Migration in that the intent is not to preserve information, but to increase or improve it. This type of AIP change may be referred to as creating a new **Edition**. The new Edition is viewed as a replacement for the previous Edition, but it may be of historical interest to retain the previous Edition.

An OAIS may also find it convenient to provide an AIP that is derived from an existing AIP. It may do this by extracting some information, or by aggregating information from multiple AIPs, to better serve Consumers. This type of resulting AIP may be referred to as a **Derived AIP**. It does not replace any of the AIPs that it was derived from and it is not a result of a Digital Migration.

6 ARCHIVE ASSOCIATIONS

Users of multiple OAIS archives may have reasons to wish for some uniformity or cooperation among them. For example, consumers of several OAIS archives may wish to have

- common finding aids to aid in locating information over several OAIS archives,
- a common Package Descriptor schema for access
- a common DIP schema for dissemination, or
- a single global access site.

Producers may wish to have

- a common SIP schema for submission to different archives, or
- a single depository for all their products.

Managers may wish to have means for

- cost reduction through sharing of expensive hardware
- enforcing submissions, or
- increasing the uniformity and quality of user interactions with the OAIS.

Therefore, it may be advantageous for OAIS archives to cooperate to meet these wishes. The motivation might come from the archives themselves, or it may be imposed by an authority that has some influence over them. In the former case, the archive might be motivated by the desire to keep consumers happy with their products, or to keep users happy with their quality of service, or simply by the need to compete with other archives in order to survive or grow. Situations like this can and have motivated agreements without the need for any explicit federation establishing an external authority. However, in cases where explicit federation is established, the external authority is represented in this Reference Model by Management.

The purpose of this chapter is to explore the degree of interaction and cooperation among archives. Section 6.1 focuses on technical levels interaction, while Section 6.2 discusses management issues concerning the tension between cooperation and autonomy.

6.1. TECHNICAL LEVELS OF INTERACTION BETWEEN OAIS ARCHIVES

OAIS associations can be categorized technically by both external and internal factors. External factors include characteristics of the Producer and Consumer communities. Internal factors could include common implementations of the information models presented in Section 4.2, or multi-archive sharing of one or more of the functional areas presented in Section 4.1.

This section defines four categories of archive association, the first three of which have successively higher degrees of interaction:

Independent - A Local community with no management interaction and no knowledge by one archive of Standards implemented at another.

Cooperating - A Local community with potential Global producers, common submission standards, and common dissemination standards, but no common access. One archive may make subscription requests for key data at the cooperating archive.

Federated - Includes both Local and Global communities, and has both Local and Global access. The Local community has priority over the Global community. Optionally, Global dissemination and Ingest are options.

Shared resources - A Local community in which Management has entered into agreements with other archives to share resources, perhaps to reduce cost. This requires various standards internal to the archive (such as ingest-storage and access-storage interface standards), but does not alter the community's view of the archive.

The remainder of this section gives a more detailed view of these categories of association.

6.1.1. INDEPENDENT ARCHIVES

An independent archive is assumed to serve only a single designated community which has participated in the requirements and design of the SIP, DIP and finding aids.. An independent archive may choose to design these structures based on formal or de-facto standards which would allow cooperation with other archives which implement the same standards. However, the design decisions to use these standards are not based on the possibility of inter-operation with other archives, but rather on local requirements and cost savings.

The classification of an archive as independent is not based on its size or distributed functionality. An independent archive may occupy one site, or may be physically distributed over many sites. It may use many standards for a given internal element. However, if there is no concept of interaction with other archives, the archive is independent.

6.1.2. COOPERATING ARCHIVES

Cooperating archives are based on standards agreements among two or more archives. Cooperating archives have related communities of interest, so they order and ingest data from other cooperating archives and possibly have common data producers. No common access, submission and dissemination standards are assumed. The only requirement for this architecture is that the cooperating group support at least one common SIP and DIP format for inter-archive requests. The control mechanism for this sort of inter-operation can be subscription requests at each archive.

The following figures illustrate the concept of cooperating archives.

At a rudimentary level of archive interaction, **Figure 6-1** represents a simple mutual information exchange agreement between archives. (Note: In this and the following figures, the OAIS is represented as a “five-port device” following the arrangement of Figure 4-1. In each case, a two-archive federation is shown for simplicity, although the concept can be extended indefinitely.) The essential requirement for this federation is a set of mutual Submission Agreements, subscriptions, and user interface standards to allow DIPs from one

archive to be ingested as SIPs by another. Therefore, it assumes that some pair-wise compatibility has been established between the archives. This does not necessarily require common access, dissemination and submission methods for all participants, although that might encourage more exchange.

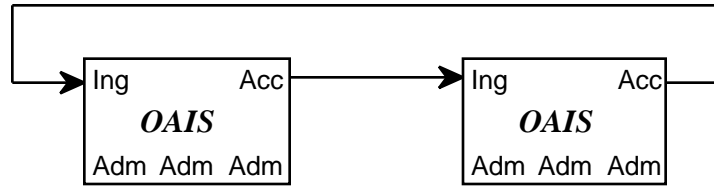


Figure 6-1. Cooperating Archives with Mutual Exchange Agreement

Figure 6-2 is an example of OAIS archives that have standardized their submission and dissemination methods for the benefit of users. No special external element is needed for this. Its disadvantage is that there is no formal mechanism for exchange of catalog information. Where does the consumer look for the desired information? Perhaps the archives agree to exchange context and catalog information, or perhaps one of the archives agrees to take on the role for both.

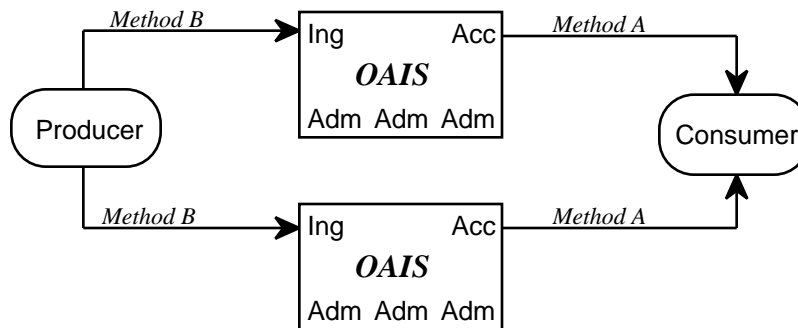


Figure 6-2. Cooperating Archives with Standard Ingest and Access Methods

6.1.3. FEDERATED ARCHIVES

Federated Archives are conceptually consumer-oriented. In addition to the local designated community of an archive, a Global community exists which has requirements for access to the holdings of the archive. However, the Local consumers are likely to have priority over the Global consumers.

At the federated level of association, external elements can be introduced to improve interoperability. For example, **Figure 6-3** shows a way to solve the catalog problem described in Section 6.1.2 using an entity external to the OAIS. Here, a pair of Producer-Consumers, each of whom maintains his own OAIS archive, has joined together to share information. The

Common Catalog & Manager is the external (global) binding element that serves as a common access point for the information in both archives. The Common Catalog may limit its activity to being a finding aid, or it may provide full combined Access service as shown in the Figure. Optionally, it may include common dissemination of products from either or both archives.

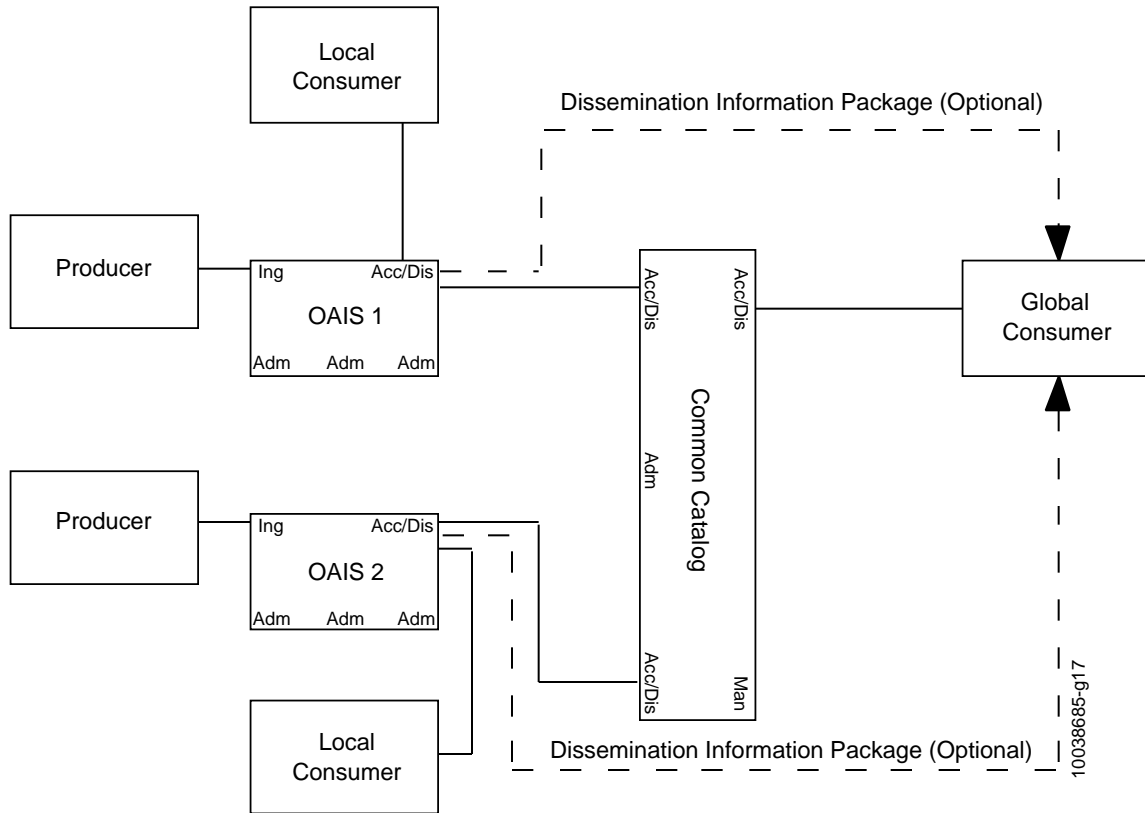


Figure 6-3. An OAIS Federation Employing a Common Catalog

The architecture of Figure 6-3 closely matches that of NASA's TIMED (Thermosphere, Mesosphere, Ionosphere Energetics and Dynamics) space mission, in which each of four principal investigators operates his own archive. As Producer, each principal investigator is responsible for archiving his own products. However, there are instances where the Producer incorporates the products of other principal investigators into his product. In this application, the mission operator is the external entity. The Common Catalog component provides an Access & Dissemination service which can assemble special products obtained from multiple archives, and can also accept subscriptions for them. Additionally, the Manager component exploits its position to enforce the submission of data products by each Principal Investigator to his own archive, according to the agreement established for the mission.

Federated archives may be further classified into three levels of functionality.

Federated 1 - Global access is accomplished by the export of a standard-format Associated Description to a global site. The global site independently manages a set of descriptors

from many archive and has finding aids to locate which archive owns a collection of interest. An example of this architecture is the Master Directory/GCMD work.

Federated 2 - Global access is accomplished by having a global node which can distribute a query to multiple local archives. This means that the local Data Management entity must store an additional Associated Description in the global format or have a translator from the global queries to local queries. An option in this case is to establish a common DIP format to ease the load on consumers who may be ordering products from many archives. Examples of this are early versions of IMS V0 or CEOS IMS.

Federated 3 - This adds to the functionality of Federated 2 a standard ordering and dissemination mechanism, available through the global node. This is a fully functional, modern federated system. Here, the global system will have some effect on the schema designs in each local archive; it would be optimal to build new local archives based on the global schemas and finding aids to ensure high degrees of inter-operability. An example of this level of inter-operability is the ECS and ICS in the earth observation domain.

6.1.4. ARCHIVES WITH SHARED FUNCTIONAL AREAS

In this type of association, Management has entered into agreements with archives to share or integrate functional areas. The motive for this may be to share expensive resources such as archive storage management or a rare peripheral device for dissemination. This association is fundamentally different from the foregoing examples, in that we can no longer ignore the internal architecture of the archive.

Figure 6-4 illustrates the sharing of a common storage function, consisting of an Archival Storage entity and a Data Management entity, between two archives, OAIS 1 and OAIS 2. The access, dissemination and ingest facilities can be at any of the previously-described levels of inter-operability. In fact, each archive can serve totally independent communities as implied in this Figure. However, it is evident that for the common storage element to work, standards are needed at the internal Ingest-storage and Access-storage interfaces. The CNES SPOT/IMAGE archive is an existing example of this architecture.

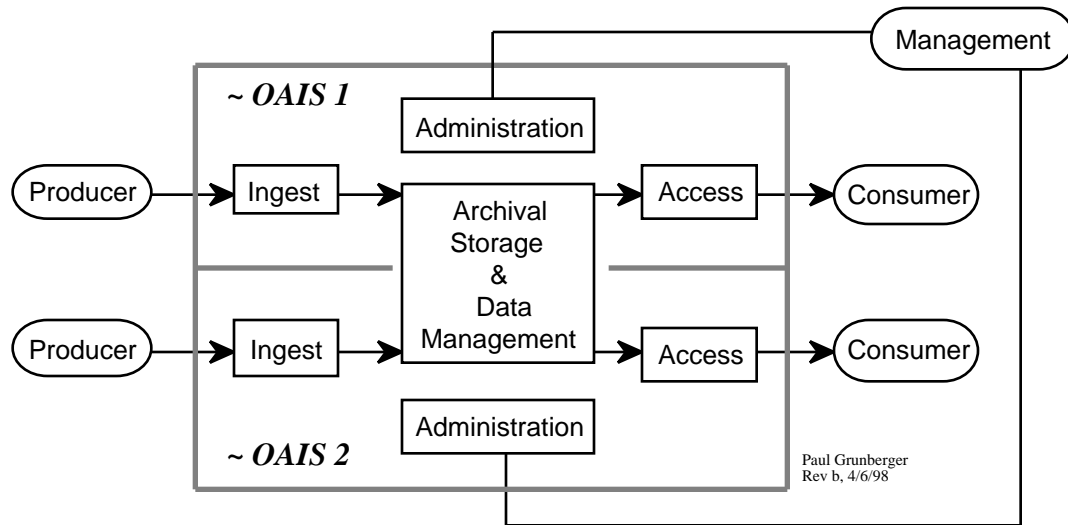


Figure 6-4. Archives with Shared Storage

6.2. MANAGEMENT ISSUES WITH FEDERATED ARCHIVES

The above examples show that the OAIS model is consistent with federation to accomplish specific objectives. However, it should also be considered that some of these objectives may be accomplished through voluntary action. This is an important dimension in association of systems, including archives, because it establishes the degree of autonomy for each system.. At the heart of the autonomy issue is the ease with which an association may be altered by one of the participants. Some possible characterizations of autonomy levels might be:

No interactions and therefore no association

Associations that maintain your autonomy. You have to do certain things to participate, but you can leave the association without notice or impact to you. An example is participation in the Internet, including operation as a domain name server. You need to meet certain requirements to participate, including the maintenance of a site with certain characteristics, but you will in effect be expunged from the association if you simply stop conforming. However, there is no penalty for this. Therefore, you have maintained full autonomy the whole time because you are free, without penalty, to do what you want.

Associations that bind you by contract. To change the nature of this association to, meet your desires, you will have to re-negotiate the contract. The amount of autonomy retained depends on how difficult it is to negotiate the changes. The difficulty may rise as more entities become a party to the contract.

So, the autonomy dimension is a key one for interacting archives, determining the ease with which each can effect changes in the nature of the association and the impact/penalty to each for recovering full autonomy. This dimension is different from the degree of technical homogeneity the association implements or supports, but it is not totally independent. For example, a high degree of technical homogeneity can be achieved in a broad association where each participating entity is free to leave without penalty. However, the permanence

of such an association can not be guaranteed, and may be improved by making it more difficult to re-negotiate the association, or by establishing penalties for recovering full autonomy. Also, a given degree of technical homogeneity might be achieved more rapidly and at lower cost when the contract is more binding.

ANNEX A. SCENARIOS OF EXISTING ARCHIVES

A.1 PLANETARY DATA SYSTEM ARCHIVE

I. DOMAIN

Domain and Consumers. The Planetary Data System is chartered to provide data archiving services, data access and expert help to the NASA-funded planetary science community. The PDS is a distributed system with a Central node at the Jet Propulsion Laboratory and discipline nodes (imaging, geosciences, atmospheres, planetary plasma interactions, small bodies, rings) located at universities around the country. The early focus has been on restoring historical mission data and has produced several hundred CD-ROM volumes containing about 80 per cent of the important planetary data archives. There has been an increased emphasis on providing access to the general public for educational outreach over the past several years.

Data Producers. Planetary data sets originate with NASA flight project data management and science teams (new data, some restorations), individual scientists (newly processed or value-added data) or via the PDS discipline nodes (restorations and value-added data). At least 50 per cent of the PDS resources have been devoted to restorations over the past seven years, with several more years of work needed to capture all historical data.

II. INGEST PROCESS AND INGEST INTERFACE

The PDS has developed a very formal interface with the major data producers (flight projects). This interface is documented in the Data Preparation Workbook and involves substantial interaction between node personnel, data engineers and project representatives. A Project Data Management Plan, signed by the PDS project manager provides the basic project data description and agreement to deliver to PDS. Since about 1993 all NASA announcements for Planetary investigations or analysis require that all data generated be delivered to PDS in conformance with PDS standards.

Submission Agreements

Projects provide a Project Data Management Plan. Sometimes a more specific document, the Archive and Transfer Plan, supplements the PDMP, providing extended product documentation and a schedule of deliveries.

Individual scientists can propose to be "data nodes" and receive funds from a PDS discipline node for preparing restored or value-added data sets for inclusion in the archive. There is no formal submission agreement for data nodes.

The PDS discipline nodes each maintain a list of outstanding restorations. These are worked-off based on their priority within discipline. At some point this list will be completed and only new project or data node data sets will be ingested into PDS. There is no formal

agreement associated with discipline data restorations.

Each data set that is identified for ingest in PDS is assigned to a Central node data engineer. It is the responsibility of the data engineer to see that all archiving steps are completed. The archiving steps are called out in the PDS Data Preparation Workbook.

Typical Data Delivery Session. Typically a delivery session will consist of a single data set contained on one or more volumes of CD-ROM or CD-Recordable media. A data set is defined within PDS to be a group of homogenous data granules at the same data level (raw, decalibrated, reduced) which differ only in time of acquisition and major category of target body. For example, the images of Jupiter taken by both Voyager spacecraft comprise a single data set. The standard process includes up-front negotiations between PDS and the provider; the production of test products which are evaluated in the peer review; revised final test products which are validated by the data engineering staff at the central node; approval and production of CD-ROM volumes; distribution by the appropriated discipline node or the central node; entry of the data set into the PDS central catalog; and entry of the data set into the NSSDC ordering system.

Transformation Process. In most cases the original data formats are maintained when data is brought into PDS. This allows existing software tools to continue to be used with the data. Much of the data preparation involves carefully documenting the data format and preparing metadata (granule labels, index files and catalog templates).

Validation. Validation is generally performed as part of the peer review of a product or by using validation tools. In some cases (for example, Magellan), the project develops its own internal validation process. The main validation tool of the PDS is the Volume Verifier. This program is run by the Central Node data engineers on each product delivered from a project or a data restoration. It validates the format and content of all product labels, and validates data files using checksums.

Security. The only area where any special security issues exist involves the receipt of proprietary data. Some projects have one-year proprietary periods before data is released to the science community. The PDS policy is to avoid receipt of any proprietary data sets during the proprietary period.

III. INTERNAL FORMS

The PDS has developed standards for documenting data sets (templates) and individual data products (PDS labels) using a keyword=value label system called the Object Description Language (ODL). Recommendations are also provided for volume organization and data product formatting to optimize the utility of resulting data products.

The PDS standards are specified in the PDS Standards Document. Standard documentation requirements include templates describing the data set, instrument, mission, etc. These templates are included on data volumes and also entered in the PDS high-level catalog. Standard terminology is maintained in the Planetary Science Data Dictionary, which is

jointly maintained by the PDS and the multi-mission ground data system. The metadata values for new data products are carefully compared with the PSDD and existing values used wherever possible. Additions are made to the PSDD to add new standard values to accommodate new data sets and when justified new keywords are added to the PSDD. Data products can have specialized metadata values which are not cataloged in the PSDD.

The PDS product labeling system is flexible enough to allow nearly any data structure to be described. Labels can be attached to the beginning of the data file or detached in a stand-alone text file which points to the data file. In some cases a single label file is used to describe multiple data files. Detached labels can be used to describe data stored in other formats (FITS or HDF, for example). In cases where complicated raw telemetry formats are stored the Software Interface Specification (SIS) for the product is included in lieu of descriptive labels.

Archive Volume Components

An archive quality data set is required to contain the following components.

AAREADME.TXT	- Text summary of data contents.
VOLDESC.SFD	- Standard volume label.
VOLINFO.TXT	- Text description of data contents.
CATALOG	- DATASET.CAT, MISSION.CAT, INST.CAT
INDEX	- ASCII index for each granule on the volume.
SOFTWARE	- Software needed to interpret/display the data.
CALIB	- Calibration data sets.
BROWSE	- Browse products for this volume.

Peer Review. All restoration and data node produced data sets are required to undergo a peer review before acceptance as archive products. Products produced by flight projects do not go through a formal peer review process. In general there is ongoing negotiation between the data engineer or the discipline node staff and the data producer. The peer review team consists of a number of scientists familiar with the data set, the discipline node leader and one or more data engineers. All product documentation and sample products and software are supplied to the peer review group for evaluation. The peer review group determines the adequacy of documentation and quality of the data products and either approves the product or provides a set of liens which must be fixed prior to approval. The PDS nodes and data engineers have access to a Volume Verifier tool which aids in validating the quality of an archive volume. The volume verifier checks internal checksums, verifies that the index contains entries for all data products and validates the volume templates as well as the descriptive keywords supplied for each product.

Delivery Media. Discipline restorations and data node products are recorded on CD- ROM or CD-recordable media as a standard practice. Flight projects are urged to provide archive quality products on CD media but may not be able to due to funding constraints. Products delivered to PDS on magnetic tape media are assigned to the PDS restoration queue. It is the

goal of PDS to convert all data sets to CD-ROM or CD-recordable media which is replicated at a separate geographic facility. This separate facility is generally the National Space Science Data Center (NSSDC) at Goddard Space Flight Center.

IV. ACCESS

Nearly all access to PDS data sets is via the CD-ROM volumes which are distributed to the entire research community. Large discipline node data collections including a substantial volume of CD-ROM data are accessible via the Internet. Several of the discipline nodes have developed on-line retrieval systems customized to meet the needs of their discipline scientists.

Finding Aids. The Pilot PDS devoted substantial resources to designing a central catalog system and distributed query and processing capabilities at the discipline nodes. These efforts were largely dropped as the Planetary Data System focused on data restoration rather than data access. In general, most of the user community already had home grown tools for data analysis and were most concerned with getting access to the data sets. The growth of the user community due to Internet and increased usage of CD-ROM readers has spurred to prototype a more consistent finding aid. The PDS Navigator has been developed for selecting images from the Clementine mission. It includes three components, a forms-based traditional database retrieval capability, an image-based retrieval and a text-based retrieval.

Security. The high-level PDS catalog can be accessed via a group account. Most of the data access services at the discipline nodes require the user to obtain a valid account on the node computer.

Customer Service/Support

The order function of the PDS is distributed. Data inventories are kept at NSSDC, the PDS central node and at each discipline node. In general each site serves a special group of users:

- discipline node - members of the NASA funded discipline
- central node - other NASA scientists and engineers, other agencies
- NSSDC - other scientists, agencies, public and foreign users.

The PDS Operator at the central node handles requests for PDS documentation or standard data products. The discipline nodes handle data requests from within their discipline and also provide expert help in the utilization and interpretation of the data. Access to tools is also provided.

V. DISSEMINATION

The vast majority of data dissemination is done via CD-ROM disc. Several hundred copies of over 500 titles have been distributed to date.

Subscriptions. Nearly all PDS distribution is done via subscriptions or standing distribution

lists. It is the responsibility of each discipline node to maintain a distribution list for its discipline scientists. This list determines the order amounts for most CD-ROM titles. The central node maintains a distribution list for engineering and management personnel and for other external recipients (reciprocal distribution, software developers).

Media/NetworkUse. Nearly all final products are delivered to the user community on CD-ROM. Archival products that need not be widely distributed are stored on CD-Recordable media, with a duplicate copy provided to the NSSDC. Most PDS data is available for downloading via anonymous ftp connection to a large CD-ROM jukeboxes at the central node and the imaging node.

Data Manipulation. Each discipline has a suite of government developed analysis tools which can be applied to the discipline data sets. These software packages are available for UNIX workstations or VAX VMS platforms. Several nodes provide the user a menu of processing functions that can be performed on selected data and will carry out requested processing and provide the results electronically or via media. The most widely used commercial tool is IDL.

Pricing Policy. The PDS distributes data to legitimate NASA researchers for no charge. There are no charges for on-line computer usage or data processing to NASA researchers. The NSSDC distributes CD-ROMs for \$10 per volume.

Security. All PDS data sets are certified GTDA by the Department of Commerce and are distributable worldwide.

VI. SPECIAL CHARACTERISTICS

PDS has invested a substantial engineering effort in its common data dictionary, data standards and procedures for preparing archival quality data sets. By having these standards in place the PDS is able to demand better quality data sets of its data providers.

A.2 NATIONAL ARCHIVES AND RECORDS ADMINISTRATION'S CENTER FOR ELECTRONIC RECORDS

I. DOMAIN

Domain and Consumers

The Center for Electronic Records is the organization within the U. S. National Archives and Records Administration (NARA) that appraises, accessions, preserves, and provides access to federal records in a format designed for computer processing. NARA serves as the archives for the records of the United States federal government. Consumers for this data are as diverse as the electronic records they seek to access and range from individuals seeking to assert their rights to other government agencies to academic researchers, private consultants, media personnel, and a wide variety of other users.

Data Producers

Originally this data is produced (created or received) by agencies of the U.S. federal government (producers). The data may concern virtually any area or subject in which the government is involved. They may come from a variety of computer application such as data processing, word processing, computer modeling, or geographic information systems. They can include records made directly by government employees or indirectly through government grants and contracts.

Special Features

The most noted special feature of NARA's Center for Electronic Records is the diversity of the collection of more than 29,000 data sets from more than 100 bureaus, departments, and other components of executive branch agencies and their contractors and from the Congress, the Courts, the Executive Office of the President, and numerous Presidential commissions. A small portion of the data originally were created as early as World War II. An even smaller portion contains information from the nineteenth century that has been converted to an electronic format. Most of the data, however, has been created since the 1960s. The major types of holdings and subject areas include agricultural data, attitudinal data, demographic data, economic and financial statistics, education data, environmental data, health and social services data, international data, and military data.

Scientific and technological data already transferred to the Center include the National Register of Scientific and Technical Personnel; the National Engineers Register; the 1971 Survey of Scientists and Engineers; major portions of the National Ocean Survey's Nautical Chart Data Base; numerous Environmental Protection Agency series relating to pesticide use, hazardous wastes, and pollution abatement; the Nuclear Regulatory Commission's Radiation Exposure Information Reporting System; biometric data sets and epidemiological studies (such as the National Collaborative Perinatal Project) from the National Institutes of Health, the Centers for Disease Control, and the National Center for Health Statistics; and text from presidential commissions on Three Mile Island, coal, and the Space Shuttle Challenger Accident. While the Center's scientific and medical holdings are rich and varied they do not fully reflect the extent and diversity of federal activity in this area.

II. INGEST

The ingest process begins with producers (records managers and records creators in federal agencies) inventorying all electronic records and determining how long to retain the records for current agency business. The next step in the process is for the producer and NARA to develop a *Request for Records Disposition Authority*, Standard Form 115 (SF 115). Here information on the content, retention and disposition, and the availability and extent of documentation and related reports is listed in the context of the producer's business needs for the information. Data with continuing value are listed as permanent and the timing and frequency of their transfer to NARA is established. The producer submits the SF 115 to NARA for its review and appraisal. The Center for Electronic Records appraises electronic

records items on all SF 115s. Identifying permanently valuable electronic records for retention by NARA's Center for Electronic Records involves cooperation between NARA and the producers. Through the process of scheduling and appraisal, the Center identifies and selects the electronic records it judges to have enduring value. The Center evaluates electronic records in terms of their evidential, legal, and informational value and their long-term research potential. Some of the factors in this appraisal evaluation include estimation of past, present, and probable future research value within the context of the data's origin and current use and its impact on federal programs and policy. Administrative and legal value, as well as the potential for linkage with other data, may bear on the decision. Unaggregated microlevel data sometimes has the greatest potential for future secondary analysis. Once the Center determines the records have enduring value, it then determines whether the records should be preserved in electronic format.

Submission Agreements

The actual Submission Information Package (SIP) between NARA and the agency that creates or receives the data is a *Request to Transfer, Approval, and Receipt of Records to the National Archives of the United States*, Standard Form 258 (SF 258) accompanied by the data object(s) and sufficient documentation and descriptive information to use the data. The SIP transfers physical and legal custody of the electronic records from the producer to NARA. This agreement is the end product of the ingest process described above. The SF 258 also contains any restrictions on access to the data which conform with exemptions listed in the U.S. Freedom of Information Act. The Center enforces all legitimate restrictions on access. The Center also works with the producer to determine if any "disclosure-free" version of the data can be produced for consumers.

Typical Delivery Session

This inventorying, scheduling, and appraisal process specifies the data object(s) and related metadata and documentation to be transferred and establishes the timing and frequency of submissions. Specific instructions for how the data are to be organized and when they should be submitted are established in the *Code of Federal Regulations* (36 CFR 1228.188). All data should be transferred on either open reel magnetic tape, tape cartridges, or CD-ROM. The CFR sets the specific technical requirements in terms of format, block size, and extraneous characters. While the current regulations also require that all SIPs should be transferred in a software-independent format, NARA staff recognize that the research potential and utility of some data would be significantly reduced if they were transferred in such a format. In such cases NARA works with the producers to determine the best mode of transfer.

What are the Information Objects that are Delivered? Producers typically will transfer a series consisting of one or more data sets with the related documentation which minimally should include the record layout and codes, methodology statements, technical information about the data including number of records and size. Ideally, the SIP also includes associated analyses and reports. Increasingly agency-created metadata also is included. The majority of electronic records come as flat files of data; increasingly, however, text files and output from

data base management systems, and geographic information systems also are transferred.

What are Collections? NARA organizes all Archival Information Collections (AIC) on the basis of Provenance and Original Order. Provenance maintains the identity of an Archival Information Package (AIP) or an AIC and preserves as much information as possible about its origins and custodial history. Within NARA this is accomplished through the use of Record Groups which reflect the structure of the federal government and subgroups and sub-subgroups which place the AIPs and AICs within the producer's place within its agency. Original order argues maintaining the contents of an AIP or AIC in the order developed and used by the producer. This helps reveal the producer's organization and how it used the data objects and can provide additional information to consumers. For electronic records, "original order" is expressed in the logical structure of files and databases and in the indexing which the producer used. Within NARA the basic unit for arrangement and description is the AIC which can include a number of related AIPs.

What Descriptive Information is Provided? The extent and quality of the descriptive information provided by the producer varies from quite sketchy to extremely detailed. NARA staff attempt to flesh out the producer-created descriptors with AIC level descriptions, title list entries, abstracts, and Dissemination Information Packages (DIP) and to provide the descriptive information in a variety of formats to reach different consumers.

What sorts of Validation Objects are Provided? Producers are required to transfer metadata and descriptors adequate to access, process, and interpret electronic records. For formatted data files the DIP must include a record layout with appropriate field definitions and codes. It frequently also includes methodology statements, input documents, data entry instructions, processing directions, sample outputs, reports and analyses of the information and system manuals.

What Transformation Processes are Performed Prior to Storage

What Metadata is Created? The most extensive metadata product created by NARA is the DIP. In the Introduction, Center staff discuss the origin, creation, and administrative uses of the data object(s), list related objects that are or will be available, and discuss characteristics of the data that could cause problems for consumers based on initial validation processes. The DIP also includes sample printouts of the data and tables and reports related to computer validation of the data. NARA also captures metadata on record layouts, domains, ranges, and links between files in a metadata database as a byproduct of the automated validation process. Other metadata created by Center staff include AIC descriptions, formatted abstracts, title line entries, and collective descriptions which place the data in a broader context. The Center anticipates that increasingly metadata created by the producer will be part of the SIP transferred to NARA.

What Validation is Performed? The Center's initial accessioning procedures include creating a new master and backup copy of each data object on new certified media to ensure the best physical media for long-term storage. At this time Center staff perform automated comparisons of the data contents with the record layout and codes, and of the physical

structure including the number of records, blocks, and bytes. Staff also perfect the DIP to facilitate secondary use of the data.

Security

All data are maintained off-line with consumer access only to copies of the data. The master and backup copies are maintained in separate secure stacks at two different physical locations. Data which require additional security measures, for example Census data subject to restrictions imposed under Title 13 of the *United States Code* and national security classified information restricted under Executive Order, are afforded the appropriate level of protection. The Center is moving to provide enhanced access to selected data onsite by providing reference copies on a wider variety of media and by providing a broader range of services and output products. This may include use of vendors who can provide enhanced access to the holdings utilizing “value-added” services.

III. INTERNAL FORMS

How do you Store your Data? All master and backup copies are stored on newly certified class 3480 magnetic tape cartridges. Some of the holdings have not yet been migrated from nine-track, 6250 bpi open-reel magnetic tape. Data are received and stored temporarily on other media including diskettes, 4mm, 8mm, CD-ROM, and various removable hard drives, although not all of these media conform with regulatory requirements.

Migration (Data). Based on recommendations from the media manufacturers, the National Technology Alliance, the National Institute of Standards and Technology, and various standards organizations, the Center has been migrating its data to new class 3480 magnetic tape cartridge when each media unit is ten-years old.

Migration (Metadata). Metadata has been stored in a variety of formats depending on the original format transferred with the data. Traditionally most metadata existed in textual format. The metadata captured in the validation process is maintained in a relational database. There are no current plans for migrating from this format, although the metadata can be exported in flat file format. The Center has been encouraging data producers to create and transfer metadata in electronic form. Within the next fiscal year the Center hopes to begin scanning and digitally converting metadata so it can be preserved and provided in an electronic format along with the data.

Migration (Format). The *Code of Federal Regulations* requires data producers to transfer all data in ASCII or EBCDIC with all extraneous characters removed from the data except record length indicators or tape marks and blocked at no higher than 32,760 bytes per block for open-reel and 37,871 bytes for class 3480 magnetic tape cartridge. When CD-ROM is used they must conform to ISO 9660 standard and the data must be in discrete files containing only the permanent data. Additional software files and temporary files may be included on the CD-ROM. The CFR also requires all electronic records to be transferred in a software-independent format. The Center works with data producers who cannot meet those requirements to determine the most appropriate transfer and storage formats.

IV. ACCESS

What Finding Aids are Provided?

Information about the holdings are available in multiple levels of detail and by multiple sources as a way to provide various consumers with information about the Center's holdings. The least specific detail is available in the 1996 three volume *Guide to Federal Records in the National Archives of the United States* where electronic records series are described in the context of the larger holdings from a producer. Other collective descriptions include *Information About Electronic Records in the National Archives for Perspective Researchers*, General Information Leaflet 37, which also is available on the Center's homepage (<http://www.nara.gov/nara/electronic>), and a title list of data sets available on the Center's homepage and as a printout. Specific electronic records series descriptions were created as formatted metadata for a portion of the Center's holdings for inclusion in a proposed automated description data base which has not been implemented. The most detailed description for any data set is the DIP. Each DIP may contain a narrative describing the data file(s), the record layout and codes for the data, a methodology, sample input forms and questionnaires, annotations regarding the data validity, and a bibliography. The Center also has established an email site (cer@nara.gov) for queries regarding the Center's holdings and services.

Security.

All of the Center's holdings are maintained in environmentally controlled closed stacks which are accessible only by Center staff. Master and backup copies of the data are stored in separate vaults in separate locations to facilitate disaster recovery. The Center's national security classified data sets are in separate environmentally controlled stacks approved for the storage of classified information. All processing is performed in limited access processing rooms at NARA or at the National Institutes of Health computer center. Computer processing is done on closed systems which require both a registered logon and personal identification number or password to access the system. Researchers do not have direct access to any accessioned data. Presently they access copies of the data that they have purchased for their own use.

Customer Service/Support.

The Center has a staff dedicated to providing reference services to the public and to the staffs of other federal agencies. The staff responds to both general and specific inquiries by telephone, letter, email, or in-person visit and fills orders for copies of specific data and their DIPs. The staff also provides information from records to respond to researcher requests such as for casualty records from the Korean and Vietnam conflicts. The staff also functions as a filter between researchers and the data producers when problems develop in understanding or interpreting the data. The staff develop a variety of informational material about the Center's holdings and services, much of which is available online.

V. DISSEMINATION

Do You Support Subscriptions?

The Center will accept a standing order (subscription) for electronic records that it receives on a regular, periodic basis from producers of the Federal government. Under current NARA regulations all subscriptions must be prepaid prior to shipment of the data.

What Media/formats do you use?

Currently the Center provides copies of data files on either nine-track open-reel magnetic tape or class 3480 magnetic tape cartridges encoded in ASCII or EBCDIC, labeled or unlabeled and written to the maximum block size requested. The Center also can provide an exact copy of records in nonstandard formats, if the agency transferred them this way, but it cannot validate or verify the contents of these files. In the past these other formats have included packed decimal, zone-decimal, binary, National Information Processing System (NIPS), Statistical Analysis Software (SAS), Statistical Package for the Social Sciences (SPSS), or OSIRSIS. The Center recently expanded its media options to include diskettes for smaller data sets and CD-ROM. On-line transfer of data remains a more distant goal.

What Transformation (Value Added) is Provided?

The Center currently preserves data as received from the producers; it does not routinely provide extracts from the data or other value-added services beyond computer validation of the data contents and enhanced documentation. Planned enhancements provide for value-added services including extracts from the data.

Pricing Policies.

The Center uses a cost-recovery fee schedule developed by the National Archives Trust Fund. Currently, the charge for an exact copy of all data on a input cartridge or reel, regardless of the number of data sets on the media is \$80.75 when copied to a class 3480 magnetic tape cartridge and \$90.00 when copied to a nine-track open reel magnetic tape or a CD-ROM. The Center charges an additional \$24.50 for each additional data set or file added to a media and \$7.50 for each subsequent magnetic tape cartridge and \$17.00 for each subsequent open-reel magnetic tape. Paper reproductions cost \$0.25 per page.

Security.

The same security considerations developed in relation to Access apply to Dissemination. The Center's national security classified data is made available only to researchers who have both the appropriate security clearances and the appropriate need-to-know. Other restricted data are made available only with prior written approval of the creating agency or under the terms of the restrictions which must be supported as a legitimate exemption under the Freedom of Information Act.

VI. SPECIAL CHARACTERISTICS

NARA's Center for Electronic Records has a diverse collection which reflects the diverse activities of the federal government. The staff shape the holdings through the process of scheduling, appraisal and accessioning. Currently, the Center acquires less than one percent of all federal records created in an electronic format. The timing of the transfer of electronic records from the creating federal agency to NARA is negotiated with the creator to ensure that the records are available for agency use for as long as necessary for current business and that they are transferred to NARA as soon as practicable to ensure their long term preservation for secondary use. NARA is the only federal agency with an explicit archival mandate for Federal records and thus the only Federal agency that preserves and provides access to a wide range of historically valuable records for the indefinite future. As such it is an archives of last resort for the electronic records of some federal agencies which undertake an active data dissemination function while there is a researcher interest in the data but whose mandate ceases or may cease once the demand wanes or ceases.

A.3 LIFE SCIENCES DATA ARCHIVE

I. DOMAIN

What is the domain and who are the customers of the Archive and who are the producers of the data? What are the special features of this archive?

The Life Sciences Data Archive (LSDA) is responsible for collecting and disseminating data of NASA funded Life Sciences space flight investigations. There are two general goals for NASA space life science research; one, to find counter measures to problems encountered by human bodies as a result of space flight, and two, to broaden the understanding of the effect of gravity on living systems. The LSDA's primary customer is the life sciences research community, but it is also used by students, educators and the general public. The data archived in the LSDA is produced by both intramural and extramural investigators funded to perform flight experiments through NASA grants. It is anticipated that the archive may grow to include data from investigations which are completely ground based.

The LSDA is a distributed archive with responsibilities distributed to LSDA Nodes at various NASA Centers and Projects with life sciences activities. There are the LSDA Project Nodes (ARC, KSC, & JSC) which are responsible for the actual collection and cataloging of data, and there is a LSDA Data Distribution Node (Central Node) which is responsible for dissemination of the data to the public.

The LSDA contains animal, plant, and human space flight data. This archive is notable in that it contains a unique collection of data describing, in considerable detail, biology experiments carried out in space by NASA over the past thirty years. The nature of the data is highly varied and spans many life science disciplines.

The LSDA is also unique in that it provides both digital and non-digital information. The non-digital data may be either reproducible or non-reproducible. Examples of reproducible,

non-digital data are video and audio tape. An example of non-reproducible, non-digital data is a biomedical sample.

II. INGEST PROCESS

Submission Agreements

There are two major types of data producers; the NASA Flight Project offices that design hardware and manage the experiment, and the NASA funded Principal Investigator.

To acquire data from the NASA Flight Project offices, the LSDA Project Nodes work closely with them to acquire data during flight operations. The LSDA assists the NASA Flight Project Offices in distributing this data to the Principal Investigators and gathering it as an archival product. As the LSDA is relatively new (1993) there is also retrospective archiving of past missions being done on a funding available basis.

To acquire data from the NASA funded Principal Investigator, there are a couple of methods of data collection currently being used depending on the “age” of the experiment. For previously flown experiments (prior to 1994) there is an informal submission agreement between the LSDA and the PI’s that is based on cooperation, and is not binding. For experiments being selected for flight (after 1994) the funding agreements include a contractual stipulation that the PRINCIPAL INVESTIGATOR must supply the LSDA with raw data, analyzed data and a final science report.

These funding agreements are finalized when proposed investigations are selected for flight. At this time the PIs are sent a letter informing them, that upon acceptance of funding they will be responsible for delivering the data collected as part of their investigation in a form usable by the sciences community one year post flight.

After a one year proprietary period, submission of data to the LSDA begins. To assist in its submission, the LSDA Project nodes send the PRINCIPAL INVESTIGATOR a Data Inventory package. The PRINCIPAL INVESTIGATOR fills out the data inventory forms and returns them to the LSDA Project Node. The Project Node then contacts the PI to begin data submission. In order to clarify the “usable form” requirement throughout the entire LSDA project, the LSDA is in the process of developing a post flight data reporting handbook which explains exactly how the data should be provided to the archive.

Typical Delivery Session

A typical submission information package (SIP) consists of two parts; 1) the Data Inventory forms, and 2) actual data. The inventory forms are considered CI and contain information about the data types (i.e. physical samples, hardcopy/photographic/video, computer files, other formats) and Data Sets (i.e. title, description, treatments, parameters measured, research subjects and Ids, date/period of collection, collection location, analysis phase, comments). The actual data consists of physical biospecimens, spreadsheets, final science reports, published articles, procedural documents, photographs, video tapes, analog tapes, digital or

printed images, and other types of digital data files (i.e. HRM).

Upon receipt by the LSDA the CI will be cataloged and Archival Supporting Information added, including; experiment and mission ID, Principal Investigator and Co-Investigators name, and other linking information.

Collections

The Archival Information Unit's (AIU) are compiled per investigation, i.e., all AIUs for a single experiment make up an Archive Information Collection (AIC). During a typical ingest session each SIP is cataloged and Archive Supporting Information is added. This CI is entered into a database comprised of LSDA approved fields and uses valid values whenever possible. The Archive Supporting Information is developed by the LSDA personnel at the LSDA Project Node responsible for obtaining the data. The Archive Supporting Information provides layers of metadata for the data collection that describe the experiment, mission, hardware, personnel, sessions, biospecimen, and research subjects from which the data was collected.

It is anticipated that future uses of the archive will involve the creation of AICs based on discipline or measured parameters.

Transformation Processes

In most cases a set of data is kept in its original submitted form. Exceptions to this case include data submitted on outdated media requiring transfer to current media. As little transformation as possible is performed on the data at ingest in order to keep costs down and to insure the integrity of the data. There are some instances where the data has been collected in an application format that is not widely available and in this case the Project Node will transform the data into a more commonly accessible format. (e.g. spreadsheets created in Supernova are translated to MS Excel).

After the LSDA Project Node enters/creates this CI or metadata for the data collection and the individual data elements, the information goes through a validation process. This post-entry validation is accomplished by a second check of the data by the LSDA Project Node Manager. Content validation is further ensured by sending the completed catalog entries to the data originator (Principal Investigator, Flight Project Offices) for verification. The Principal Investigator reviews the information, makes corrections or additions and sends the information back to the Project Node. Edits are then made to the records and the information is once again printed and sent to the Principal Investigator. This process is repeated until the Principal Investigator is satisfied that his experiment data is accurately represented. At this point the Principal Investigator signs and returns a verification letter to the Project Node. The catalog or metadata is now ready for review by the LSDA Project Scientist before it is placed in the public record (via Web Site). The LSDA Project Scientist will review the data for overall form and cogency. Do you want to put anything in here about the LSDA review cycle or is review by the LSDA Project Scientist a sort of catch all term?

Security

The LSDA has strict security measures for data from human subjects which require sensitivity and secure handling due to the Human Data Privacy Act. Human data when received is coded to protect the identity of the crew members. Security procedures include keeping the data on magneto-optical disks stored in a locked file cabinet in a cipher locked room.

Overall security procedures stipulate that all digital data are backed up on a daily basis with off-site storage. Access to on-line servers is controlled through the use of password and/or address port filtering. Only data that is fully validated and approved for release is placed on publicly accessible servers.

III. INTERNAL FORMS

Storage

The LSDA back-up and storage procedures vary between LSDA Node types. Currently, the LSDA Data Distribution Node resides at Johnson Space Center. The LSDA Master Catalog and on-line data reside on a Microsoft SQL Server. These are backed up to tape daily. At the LSDA Project Nodes most of LSDA's data and metadata are stored on magnetic disks and backed up to tape. Long term storage is provided on CD-ROM. Biospecimens are stored in -80 degree freezers.

AIUs are stored as a piece of CI (a spreadsheet, word processing document, strip chart or biospecimen) with the archival preservation information stored in a database record. The CI can only be, easily, accessed through the database record with it's descriptive information and the CI storage directory information. These AIUs are linked, through the database, into AICs via an Experiment Number. A space life sciences experiment is, in this sense, an AIC. It is a collection of tens or hundreds of AIUs.

Migration

The LSDA migration process is still in a developmental phase but there is some ongoing data migration. LSDA Project Nodes are in the process of converting information on outdated media (RA60's, RL02's) to CD-ROM format.

Migration of application formats (e.g. MS-Excel) and in particular, version changes, is an area of concern. The cost of continually updating all LSDA spreadsheets to the current version is prohibitive and storing and making available the application is also expensive and complicated. A universal read only format such as Adobe Acrobat might be the solution, but it is a proprietary format and it's life span is an unknown.

IV. ACCESS

Finding Aids

Access to LSDA information and data is handled through the LSDA Data Distribution Node at JSC. Users enter the LSDA through the World Wide Web (WWW) and search/retrieve information via the Master Catalog. The Master Catalog is a relational database with a WWW forms interface and allows users to search Archive Supporting Information across experiments and Missions to find data that meets their search criteria. Users can search within ten information groups; Experiments, Missions, Data Sets, Hardware, Documents, Personnel, Specimen or Subjects, Data Collection Sessions, Biospecimens, and Images.

There is currently no method available for searching data at a 'sub-AIU' level. The AIU record contains a considerable amount of detailed, searchable, data so that a collection of AIUs could be found for a particular manual sub-search.

Security. The LSDA does not have any special security concerns for access to the public Master Catalog information and non-human digital data. It is freely available to anyone on the WWW. However, the human flight experiment data is subject to the Human Data Privacy Act, and therefore, security measures are required to control access to this data. The policies and procedures for access to the data are currently being developed.

Customer Service/Support. The LSDA provides user support for questions and problems concerning the Master Catalog (on-line data request system) and for questions about the data being provided. The primary means of user feedback and support is through the LSDA Data Distribution Node. Questions are addressed to the LSDA through on-line "What do you think?" links located throughout the system. From these links a WWW forms interface allows users to submit questions. Specific questions about the data are currently addressed by the NASA Life Sciences Acquisition Scientist and the LSDA Program Scientist. Questions which can not be answered by these individuals are forwarded to the LSDA Project Node which collects the data. In some instances questions are forwarded to the Principal Investigator or NASA Flight Project Office who provided the data.

V. DISSEMINATION

Subscriptions. The LSDA does not support subscriptions since the publicly available Master Catalog is accessible to all users. The LSDA data is located using a catalog on the WWW. Most data is disseminated to the user through links in the catalog to an anonymous FTP site from which the data is downloaded. This means of data dissemination is, therefore, tightly linked to the data "finding" process. If data are in non-digital format, but are reproducible (i.e., hardcopy documents, or log books), users may request them through on-line ordering forms available in the Master Catalog. The requested information is reproduced via photocopying and shipped US Mail to the requester.

There are discussions about an update notification service to be offered in the future.

Media/Formats. The LSDA contains unique non-reproducible pieces of data such as microscope slides and space flight biospecimens. These unique resources are provided to a requester after a scientific proposal has successfully undergone peer review. Biospecimens,

once disseminated, are used to produce original data which is then ingested into the archive.

Transformations. Currently, the LSDA does not provide many value added services. The data is stored and disseminated as provided by the data producer. Data are available in raw and summarized form. These summarized data are provided by the data producer. LSDA does convert data which are received in a non-standard format to a more usable form. Currently there are no data analysis tools available through the LSDA. However, the LSDA Project Nodes do ensure that all data sets have the minimum amount of information needed for understanding (e.g. explanation of all column headings are provided for the spreadsheets, etc.). These are the only “value added” processes that come from the raw data.

Security. Since the Master Catalog and non-human data in the LSDA are available to anyone on the WWW there is no special security in place for the dissemination of this data. However; the human flight experiment data is subject to the Human Data Privacy Act, and therefore is not openly available to users. Limited dissemination of human data will be allowed using policies and procedures that are currently being developed.

Pricing Policies. LSDA data that has been verified and cleared for release is available to the public, free of cost, through the Internet. If significant requests are generated for hardcopy documents, a processing fee for copying the document may be charged. As yet this has not been determined. In the future CD-ROMs with data may be generated. These CDs will be priced in order to recoup production and distribution costs.

A.4 NATIONAL COLLABORATIVE PERINATEL PROJECT (NCP) 1959-1974

I. DOMAIN

Domain and Customers

The National Collaborative Perinatal Project was a multi-institutional, multi-year study of pregnant women and the children born from those pregnancies to provide baseline information useful for later determining the causes of neurological diseases which appeared in a portion of the studied population. The data came from medical histories, examinations, and observations. The records also contain socioeconomic, family history, and family health information. The data are used by a variety of medical and other researchers.

Data Producers

The predecessor to the U.S. National Institutes of Health’s National Institute of Neurological Disorders and Strokes (NINDS) began the National Collaborative Perinatal Project in 1958. Fourteen university-affiliated medical centers across the United States participated in the study. Between 1959 and 1965 each cooperating medical center collected information on between 300 and 2000 pregnancies each year for a total of 55,908 pregnant women utilizing their clinic services. This represented between 14% and 100% of the women utilizing these services depending on the sampling rate employed at each clinic. The final population was reduced to 39,215 due to miscarriages prior to twenty weeks, 445 multiple births, exclusion

of subsequent or repeat pregnancies, and deletion of incomplete records due to women withdrawing from the study prior to its completion. The children were given neonatal examinations and follow-up examinations were through eight years of age. The last examinations were conducted in 1974. The computer data files resulting from the research that NINDS transferred to NARA consist of approximately 6,200,000 records organized into a Master File, a variable file, and eighteen work files, one of which consists of thirteen distinct data files.

Special Features

The Collaborative Perinatal Project was a longitudinal multi-disciplinary research effort which sought to relate the events, conditions, and abnormalities of pregnancy, labor, and delivery to the neurological and mental status of the children of these pregnancies and their siblings through eight years of age. The study sought to link any later appearance of cerebral palsy, mental retardation, learning disorders, congenital malfunctions, minimal brain dysfunction, convulsive disorders, visual abnormality, or communicative disorders to patterns during the perinatal period in order to develop strategies for prevention and intervention. The sample population is large enough so that statistically significant numbers of such disorders would appear in the children. Study of the records relating those children could result in the development of predictive factors and possible preventive care or intervention actions which could reduce future incidence rates.

The data are available in two formats: microfilm of the individual case files for the mother and child of approximately 270 pages per case file and the computer data files. Access to the microfilm and two of the computer data files is restricted because they contain personal identifiers. The National Archives has created a public use file for the Master File and Work File 16: Serum Specimen Inventory.

II. INGEST

The ingest process for transferring any federal agency records to NARA begins with the agency identifying the records and assessing their potential evidential, legal or research value. The next step is for the agency to develop a Standard Form 115, *Request for Records Disposition Authority*, and submit it to NARA. NARA staff then appraise the records in terms of their evidential, legal, and informational value and their long-term research potential. NARA and the creator then establish a transfer date, negotiate any restrictions on access, and initiate the ingest process.

Ingest for the NCPP computer data was a two phase process. In Phase One, from 1958 through 1974, NIH's NINDS funded the project and the cooperating institutions conducted the research. Contractors accumulated the original examination records, created the consolidated case files, microfilmed the records, normalized the data, and developed the Master File, an extract file of frequently used variables, and special files such as "refined diagnoses". The data were stored on 23 reels of magnetic computer tape recorded at 1600 bpi. Prior to 1980 the data were available only to NINDS, the cooperating hospitals, and selected government researchers.

In Phase Two NINDS developed the documentation necessary for more generalized use of the data and negotiated a submission agreement, including access provisions, with NARA. Since the nonreleasable data could be made anonymous through creation of a Public Use File, the producer and NARA worked on transferring the data files first.

Submission Agreements

NARA and NIH executed the U.S. Government's standard transfer form, Standard Form 258, *Request to Transfer, Approval, and Receipt of Records to the National Archives of the United States*, in mid-1985. This transferred legal custody and preservation responsibility to NARA. A similar agreement for the microfilm examination records was executed in 1990 after NARA and NIH resolved the privacy and access concerns and NARA developed a statistical research form.

Delivery Session

The delivery session was a single transaction in which NIH provided NARA with copies of the 23 reels of magnetic tape containing the NCPP and the related documentation consisting of seven volumes containing the background of the study, the sample, data collection and data processing overviews, record layouts and coding for each variable, sample forms, and a bibliography of all published research through 1985. NINDS transferred the 8000 rolls of microfilm containing the examination records in 1990.

Transformation Process

NARA has maintained and preserved the original data format. The data are in a hardware and software independent EBCDIC format which facilitates wide researcher access. All data were copied to new nine-track open-reel magnetic tape when received in 1985 and are migrated to new media every ten years to ensure long-term preservation. The more than 7000 pages of documentation are available in both a page format and in a microfiche format on 75 fiche. The documentation has not been scanned or digitized.

Validation

Validation was performed as part of quality control throughout the life of the NCPP. Extensive use of the data during the life of the project (1958-1974) and its use by NIH approved researchers (1958-1985) provided a second de facto validation. NARA also validated sample portions of the data at the time of ingest. Continuing researcher use also validates the data contents.

Security

The computer data is maintained in environmentally-controlled closed stacks which are accessible only to Center staff. Master and backup copies of the data are stored in separate

vaults in separate locations to facilitate disaster recovery. NARA has created Public Use Files of the restricted data files to prevent unauthorized access to personal and medical data.

What Descriptive Information is Provided?

NARA has prepared multiple levels of descriptive information for the NCPP. These range from entries for each data file in the Center for Electronic Records' Title List, an abstract entry for the series, a series description, a full documentation package, to a series-level entry in the three-volume *Guide to Federal Records in the National Archives*.

What Validation Objects are Provided?

During its active life the NCPP established and used elaborate data collection, input and verification procedures. Extensive use also validates the information. NARA's routine transfer and storage procedures also validated the data. The extensive seven-volume documentation includes record layouts and codes, methodology statements, data analyses, and a bibliography of research use.

What Transformation Processes are Performed Prior to Storage?

What Metadata is Created? NARA staff supplemented the documentation with an abstract and introduction discussing the origin, creation, and uses of the data, including an explanation of restrictions on access and the characteristics of the Public Use File.

What Validation is Performed? NARA's accessioning and storage procedures included creating a new master and backup copy on new certified magnetic media and creating a Public Use File of the two restricted data files. Sample portions of each data set also were verified against the documentation.

III. INTERNAL FORMS

Storage. NARA maintains separate sets of the master and backup copies of the data and the Public Use Files on newly certified 3480 class magnetic tape cartridges.

Migration (Data). NARA migrated the NCPP from 23 nine-track, 1600 bpi open-reel magnetic tapes it received in 1985 and stored the data on seven nine-track, 6250 bpi open-reel magnetic tape. NARA migrated the data to four 3480 class magnetic cartridge when the media was ten years old.

Migration (Metadata). NCPP metadata is available in textual (7000+ pages) and microfiche (75 fiche) forms. There are no plans to scan or digitize the text.

Migration (Format). The data currently are encoded in EBCDIC with all extraneous characters removed. There are no plans to migrate the format at this time.

IV. ACCESS

What Finding Aids are Provided?

Information (of varying detail) about the NCPP is available in the 1996 three volume *Guide to Federal Records in the National Archives of the United States*, where the records are described in the context of the larger holdings of the National Institutes of Health; in *Information About the Electronic Records in the National Archives for Perspective Researchers*, General Information Leaflet 37; in *Title List: A Preliminary and Partial Listing of Data Files in the National Archives and Records Administration*; and in the documentation package for NCPP. Much of this information is available on the Center's homepage (<http://www.nara.gov/nara/electronic>) or by posting an enquiry to the Center's e-mail site (cer@nara.gov).

Security

NCPP, like all of NARA's holdings, is maintained in environmentally-controlled closed stacks which are accessible only by authorized Center staff. Master and backup copies of the data are stored in separate vaults in separate locations to facilitate disaster recovery. Researchers do not have direct access to the data. Presently they acquire copies of the data on a cost-recovery basis permitting indefinite use of the data for their own purposes.

Customer Service/Support

The Center has a staff dedicated to providing reference services to the public and to the staffs of other federal agencies. The reference staff responds to inquiries by telephone, mail correspondence, e-mail, or in-person visits. They fill orders for copies of all or part NCPP and the relevant documentation. The staff also function as a filter between researchers and the NINDS when problems develop in understanding or interpreting the data.

V. DISSEMINATION

Do You Support Subscriptions?

NARA's Trust Fund is willing to establish accounts that allow researchers to acquire data that is transferred on a recurring basis. Since the NCPP stopped collection data in 1974 there is no need for a subscription for this data.

What Media/Format do you use?

Copies of the 32 data sets comprising the NCPP are available on seven nine-track open-reel magnetic tapes, six 3480 class magnetic tape cartridges, or two CD-ROM.

What Transformation (Value Added) is Provided?

The NCPP is provided as received from NINDS. NARA has created Public Use Files for the two data files containing personal identifiers in conformance with the Freedom of

Information Act and NARA restrictions on access to records whose release might result in unwarranted invasion of personal privacy.

Pricing Policies.

Electronic data sets are available on a cost-recovery fee schedule developed by the National Archives Trust Fund. Currently the charge for an exact copy of all NCPP data on a storage reel or 3480 class cartridge is \$80.75 when copied to a 3480 class magnetic tape cartridge and \$90.00 when copied to a nine-track open-reel magnetic tape. Copies on CD-ROM are \$90.00 for the first file and \$24.50 for each additional file written to the CD-ROM.. Paper reproductions cost \$10.00 for the first 20 pages and \$5.00 for each additional block of 20 pages. Microfiche reproductions cost \$2.10 per fiche.

VI. SPECIAL CHARACTERISTICS

The National Collaborative Perinatal Project was a prospective study. NINDS expended more than \$200 million over two decades to collect information on more than 58,000 pregnant women and their children at fourteen cooperating institutions. It is unlikely that a study of this duration and magnitude will be repeated. The data continue to constitute an important resource for biomedical and behavioral research in many areas of obstetrics, perinatology, pediatrics, developmental psychology and other fields.

A.5 ARCHIVE SCENARIO FOR THE *CENTRE DES DONNEES DE LA PHYSIQUE DES PLASMAS* (CDPP)

I. DOMAIN AND CUSTOMERS

The CDPP (*Centre des Données de la Physique des Plasmas* - Center for Data on Plasma Physics) is a new service currently being set up. It has been developed to ensure the Long-term conservation and availability of natural Plasma Physics data (magnetospheric plasma, planetary plasma etc.) for the international scientific community. More specifically, the data concerned is from either ground-based or space-flown experiments in which France has participated or wholly directed. The CDPP is designed around two principal components:

- A Technical Activity segment, located on the premises of the French space agency, CNES, mainly in charge of developing and maintaining the archive system. The latter has the following functions: addition of data and metadata to the system, preservation of data and metadata, organization of search and product ordering facilities, and dissemination.
- A Scientific Activity segment, located at the CESR (*Centre d'Etudes Spatiales des Rayonnements* - Center for the Study of Space Radiation), a science laboratory near CNES. The CESR is in charge of all aspects relating to scientific knowledge of the data: validating data with its producers, ensuring that the data is useable by the scientific community, setting up added-value services etc. This Center is also responsible for developing a WWW server to present CDPP services, supplying educational information on Plasma Physics to the general public, and guiding users to access and dissemination functions.

The two complementary segments work closely together.

A number of associated laboratories will be able to join the two main components of the CDPP provided they offer a service (data dissemination or information) relating to natural plasma physics.

The archive system is currently being developed. The service is planned to be made available to the scientific community on 1/1/99.

Data Producers. Data producers are mainly either current or future experiments, or projects concerned with rehabilitating existing data. Ongoing experiments include, for example, the French experiments flown aboard Russian satellites (INTERBALL), aboard the US satellite (WIND), aboard the future European satellites (CLUSTER), or even some data from the EISCAT radars. The projects to rehabilitate existing data cover a many French experiments performed since 1975, mostly flown on European, US and Soviet satellites or probes.

II. INGEST

The CDPP has drawn up a specification for deliverable data products. The specification defines the characteristics (either mandatory or optional) that the data and metadata to be delivered to the CDPP must exhibit. It defines the rules systematically applied with respect to:

- file structure, data encoding and standardization of times and dates,
- orbit or trajectory data,
- the minimum content and format of catalogues,
- complementary information needed to use or interpret the data,
- etc.

The CDPP provides technical support in order to apply this specification to each data-producing project.

As far as future projects are concerned, the authorities empowered to make decisions on projects will make the drawing up of an obligatory data management plan. The plan must define exactly which data will be archived (physical values, raw data...), how the data will be organized, and when the data will be delivered to the CDPP.

One particular service within the CDPP is the SPID (*Service de Préparation des Informations et des Données* - Service for Preparing Information and Data), in charge of the interfaces with data-producing projects and the formatting of some metadata before its delivery to the archive system.

Submission Agreements. As far as future projects are concerned, the submission agreement shall be constituted by the project Data Management Plan, to be approved by both the project

and the CDPP. As far as existing data to be rehabilitated is concerned, the framework is less formal: there is normally no project team left and no longer a budget specific to that project. Rehabilitation is thus the responsibility of a team of engineers from CNES and those of the Principal Investigator or members of his team. The CDPP suggests priorities for the work to be completed. It also influences the choices and compromises to be made with regard to the level of data to be archived.

Delivery Session

Data delivery. Data-producing projects must normally store the data produced before delivery. They do so using the facilities offered by the STAF, a multi-mission storage service at CNES. The main function of the STAF (*Service de Transfert et d'Archivage des Fichiers* - Service for Transferring and Archiving Files) is the Long-term physical storage of information. The interface is stable and therefore the technologies and storage media can thus be replaced or changed in-house without affecting the interface. The STAF also monitors and renews the media used.

From a user project viewpoint, the STAF appears as a virtual tree structure in which files may be stored. When all the data to be delivered has been produced, the delivery process merely amounts to a change of ownership of the STAF directories in which the data is stored. There is no actual physical movement of data.

Delivery of metadata. Metadata generally takes up less space than data. A delivery disk space is set up by the CDPP and the data-producing project has the right onto write to this space. When all the data and metadata has been delivered, the SPID can begin its checking and formatting (see below). This process is valid for a complete set of data, a partial delivery or an update of previously delivered metadata.

Transformation Process

The format of experiment data is not altered during the delivery process. On the other hand, metadata delivered will be subject to a kind of packing (without changing the contents) and new metadata will be created by the SPID. To give some examples:

The archive system manages the descriptions of both data collections and objects, browse data and documentary information in the form of graphs on collections and objects. The delivery of a new collection results in the creation of a new node in the data description graph and logical links with existing collections. The creation of this information, granting a global and consistent view of all the data and metadata available, is not within the domain of the data producer.

When the Principal Investigator delivers a Microsoft Word document describing an experiment, he places the corresponding file in the delivery disk space. The SPID will then use this file to create a documentary object descriptor giving the document title, author, publishing body, language, associated keywords, stating the existence of an abstract etc.

The insertion of metadata in the archive system is mostly based on use of PVL (Parameter Value Language) and a DED (Data Entity Dictionary) which is configuration managed. One of the roles of the SPID will thus be to create this new metadata and construct the PVL structure describing it. Generally speaking, metadata appears as an extremely heterogeneous set of information objects. Using PVL means that these heterogeneous objects may be delivered in both a homogeneous and standard format.

Validation

The SPID is responsible for ensuring that the deliverable product specifications for each data set have been respected. It also performs a number of coherence checks, such as checking coherence between catalogue data and the files containing experiment data.

Once these checks have been completed, the results, together with all the metadata, are presented at a formal peer review whose purpose is to decide whether the CDPP can accept the data set and issue recommendations in this field. Once accepted, the CDPP becomes the guarantor of the data set. This review brings in scientists from outside both the CDPP and the Principal Investigator team.

Despite the various checks carried out, the scientific validity of the experiment data delivered to the CDPP remains the responsibility of the Principal Investigator or data-producing project.

Security. The delivery process for both data and metadata takes place within a dedicated environment accessible only by the data producer and the CDPP.

III. INTERNAL FORMS

Storage. The STAF multi-mission storage service (see above) takes charge of the data and metadata. This service currently uses StorageTek silos with 3490 cartridges and larger capacity Reedwood cartridges (10 Gigabytes uncompressed). The objects archived by this service are files. There are several different layers of service with regard to file retrieval time and file duplication. The STAF is in charge of all data migration involved when changing from old to new media or to a new technology medium. They do not affect the upper layers of the system.

Formats. The format of data stored must be independent of all operating systems. In practice, experiment data is usually in IEEE or ASCII code and divided up into sequential files. The application of CCSDS encoding for times and dates is compulsory for all record structure files. The syntax and semantics of each file must be described with EAST and a DED unless self-descriptive structures such as FITS or NCAR are used. As far as documentary information is concerned, no reference standard for the internal representation of documents has yet been applied.

Data Management

Data management revolves around use of a graph describing data collections and objects. For the purposes of simplification, this graph is usually known as a data graph. It is oriented and non-cyclical. The relations associating a node with its descending nodes are (from an object-oriented viewpoint) inheritance and composition relations. A data set, also known as a terminal collection, thus inherits the characteristics of all the collections above it.

Documentary information, browse data and event tables are also managed through graphs which are nonetheless distinct from the data graph. The graphs contain either explicit metadata or references to external files or documents.

IV. ACCESS

Access facilities are seen by the user through a WWW server. These facilities include aids to search for data collections and objects, means of retrieving certain metadata (such as documents and catalogues) immediately and ways of ordering data products which include special protective mechanisms for data not made public.

Finding Aids

The aids to search data of interest to the user are based on navigation within the different graphs: the experiment data collection and object graph, the browse data graph, the documentary object graph and the events table graph. These graphs are independent but a certain number of links are used to move from one to another. Navigation within the graphs is, depending on the case, through criteria such as a keyword (parameter measured, location of measurements etc.), time or other types of criteria.

The data object and collection graph grants several views of the data, and the final objects may be selected after several navigations within the graph.

The events table graph may be used to make indirect selections over time, such as selecting only data corresponding to a given instrument operating mode, or data corresponding to the periods during which a particular type of magnetospheric event was observed etc.

These aids may be used to select data which is stored either on the main archive site (at CNES) or at an associated laboratory.

Security

Without exception, metadata is visible and accessible to the general public without any prior authentication. On the other hand, data may only be ordered by a user previously authorized by the CDPP, as it normally implies the consumption of resources. The user makes his request for authorization by a form available on-line, indicating his name, e-mail address, the name of the laboratory he belongs to and the reasons for his request. Once the user has received authorization to order products, he must authenticate his request (name and password) before ordering.

Data archived by the CDPP is usually public in nature, but in the case of recent data, data ordering may be temporarily restricted to one particular user group. The system must therefore be capable of handling access rights to the service (for ordering data) independently from access rights to the data itself.

Finally, the system is designed and has a number of protective measures such that any accidental or deliberate modifications to the data stored in the Center may be avoided.

Customer Service/Support

The system can handle profiles peculiar to each user, taking into account in particular the capability of the network linking him to Internet and the laboratory to which he belongs (laboratories directly supported by CNES, laboratories involved in cooperative projects with French laboratories, other laboratories etc.).

The CDPP has a customer support team able to reply to technical questions (how to use the system, read data etc.). This team can also direct the users to the Principal Investigator or data producers.

V. DISSEMINATION

Subscriptions. In its initial version, the system only accepts orders relating to data available in the system.

Media/Network Use

The data from Plasma Physics experiments is often bulky (a data set often contains between ten and several hundred Gigabytes). It is not planned to systematically create pre-defined, widely disseminated products as is often the case for planetary data, particularly as users are often interested in a specific period of time and not the whole data set.

Products may be delivered either over a network or on a variety of media (currently CD-ROM, DAT or Exabytes). The choice between these two types of delivery depends on the capacity of the network between the user and the CDPP at any given time.

As far as network deliveries are concerned, the system proposes the HTTP protocol at the user's initiative or the FTP protocol at the CDPP's initiative, but at a time specified by the user. The latter choice is subject to certain constraints. Deliveries of data via a network offer optional data compression and grouping facilities in the form of .tar files.

Data Transformation

The data objects distributed to scientific users are not necessarily identical to the data objects stored in the system. Depending on the standards respected and tools available, a certain number of transformations of archived objects may be requested, in particular:

- Time-related retrieval which provides data corresponding to one (or more) time periods specified by the user. This kind of retrieval is only possible when times and dates have been encoded in compliance with CCSDS recommendations.
- Retrieval of fields, which permits the user to select fields of interest on the basis of an EAST data descriptor.

These transformations are known as "subsetting services". Other such transformations are planned for future versions, so as (for example) to be able to deliver data in the user's native machine format, or deliver data as physical values although it is stored as raw values.

Pricing Policy. The pricing policy has not yet been determined, but will probably include an invoice for dissemination of data on an external medium (CD-ROM, DAT, Exabytes).

Security. Whether data is public or not, data products ordered by a user are only visible and accessible by that user, whatever the mode of delivery.

ANNEX B. COMPATIBILITY WITH OTHER STANDARDS

(TBD)

ANNEX C. BRIEF GUIDE TO THE OMT

A key to object relationships in the OMT diagrams of this document is shown in Figure E-1.

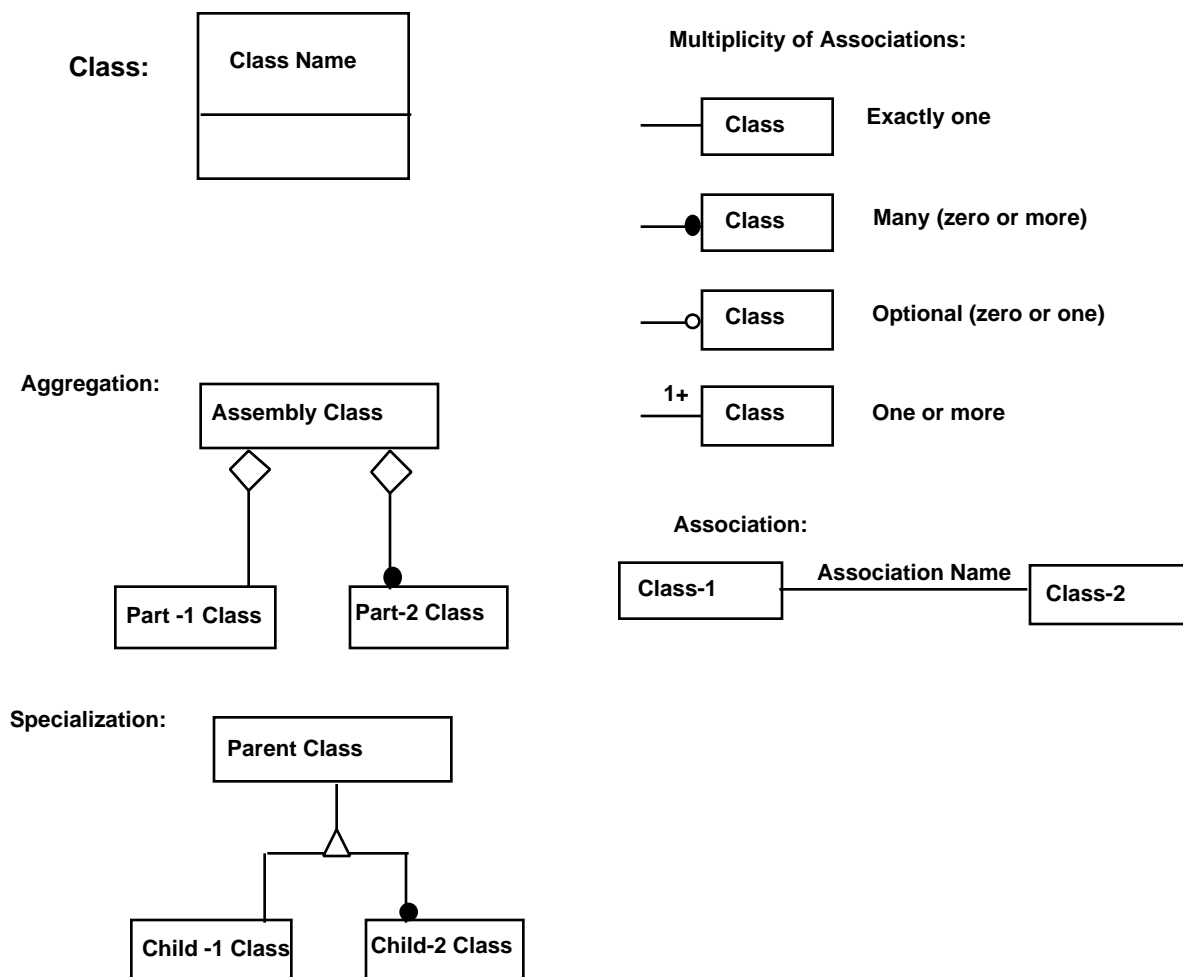


Figure E-1. Key to OMT relationships

A Class is indicated by a box, as shown, and generally there will be many possible instances of the class. Classes of objects are related to one another through Associations and there are various multiplicities that may be attached to these associations as shown. The multiplicity refers to the number of instances, or objects, of that class that are involved in the relationship.

The general association among two classes is given by a connecting line that is labeled with an information phrase indicating the nature of the association. There are two particular associations that are commonly used - aggregation and specialization, and these have particular symbols to indicate them.

An Aggregation association is one where a class is considered to be a part of another class.

A Specialization association is one where a child class inherits properties from the parent class. A child can do what the parent can do, but generally can also do more.

In the figure, the aggregation association says that the Assembly class contains exactly one Part-1 class instance and zero or more Part-2 class instances. The specialization association says that the Parent class properties are inherited by one instance of the Child-1 class and by zero or more instances of the Child-2 class.

ANNEX D. ENTITY AND FUNCTION MATRIX

Table D-1 shows the functions identified under each major entity in Section 4.1, and aligns analogous functions in rows.

Table D-1.
Sub-Functions of the OAIS Entities, Aligning Similar Functions in Rows

Entity Function Category	Ingest	Archival Storage	Data Management	Administration	Access and Dissemination
PLANNING	Schedule Submission Delivery	Manage Storage Hierarchy	Administer Data base	Negotiate Submission Agreement	Prepare Finding Aids
INPUT	Receive Submission	Receive AIPs	Receive Data base Updates	Receive Status	Receive Requests
QUALITY ASSURANCE		Error Checking		Audit AIPs	
PROCESSES	Generate AIP	Refresh Media	Perform Queries	Develop Standards and Policies	Coordinate Requests
	Generate Descriptive Information		Activate Requests	Manage System Configuration	Generate DIPs
OUTPUT	Coordinate Updates	Provide Data	Generate Reports	Interact with Management	Deliver DIPs
SECURITY		Provide Disaster Recovery		Provide Physical Access Control	Provide Access Controls

ANNEX D. ILLUSTRATIVE SCENARIOS

The scenario in this section describes the flow of information into and out of a hypothetical archive and describes the organization and maintenance of the information within the archive. The scenario is based upon a specific example — an archive containing the medical records of a hospital — but the same principles found here apply to a wide variety of applications, including: archival of the records of a government agency; or the financial files of a multinational corporation; or the computer-aided design drawings for a company's products; or scientific data on a particular topic gathered by researchers from around the world. For all these applications the intent is to preserve information for an extended period — decades at least, even centuries in some cases.

We assume that the archive is already established and operational, with hardware, software, policies, procedures and personnel in place. We begin therefore with the preparations for placing a new kind of information into the archive. The hospital decides to archive the diagnostic images from a Computer Tomography scanner. The producers — members of the hospital's radiology department — meet with the archivists to pave the way for this submission. The archivists are not experts in CT imagery, so during this discussion the radiologists describe the data they will be submitting and how they expect the data will be used in the future. Together the producers and archivists draw up a *Submission Agreement* that spells out the form in which the data will be submitted, the frequency with which it will be submitted, the length of time for which the data are to be archived, the disposition of the data when that time expires, limitations on who may access the archived data, and so on. The submission agreement defines the content and format of the *Submission Information Packages* that will be provided to the archive. The SIP will include any or all of the following: (1) content information; (2) preservation information; and (3) packaging information.

Content information consists of *data objects* and *representation information*. Data objects are the most important part of the submission and the other parts of the SIP are provided chiefly to ensure the proper archival and eventual retrieval of the data objects. Data objects can be either *physical objects* or *digital objects*. An example of a physical object would be an X-ray image recorded on film; a digital object would be a 3-dimensional CT image recorded on a computer-readable medium. *Representation information* is the next most important part of the submission: it is information that is essential to understanding a data object but that is not provided as part of the data object itself, and which will likely not be available to users through outside sources. For either an X-ray or CT scan this might include some information on the patient — name, sex, age, and so forth — as well as information on the X-ray or CT device — brand and model, exposure settings, etc. A digital object like a CT scan will furthermore require some representation information that is not needed for a physical object: namely, a description of how the bits and bytes that make up the data object are arranged. For example, this kind of representation information would specify that the CT images are composed of pixels, each of which is an 8-bit unsigned integer number; that pixels are arranged in a certain order into image planes of a specific size, and that a specific number of image planes (in a certain order) form the 3-dimensional CT scan.

Preservation Information is additional information required to maintain the content information within the archive and to ensure that the content information can be interpreted and used by consumers at some time far in the future. Preservation Information comprises four kinds of information: Context; Provenance; Reference; and Fixity. Any or all of these may be included in a SIP. They may also be added by archivists when the SIP is ingested. Here are some examples of these kinds of information:

- *Context* — Information linking content information to a larger environment. For the images from our CT scanner, context information might include the ID number assigned to the patient by the hospital, which can be used to link the images to the patient's other records. Context information might also include a reference to a technical description of the CT scanner that might be needed by future doctors to properly interpret the images.
- *Provenance* — Information describing how the content information was processed or handled before being inserted into the archive. Take, for example, the data from a scientific instrument. The data submitted for archival have been calibrated to convert the unique output of the instrument into a physical quantity (temperature, pressure, etc.) that can be compared with other scientific data. The calibration that was used in this process would be included in the submission as provenance information.
- *Reference* — A reference is an identifier by which the content information is known outside of the archive. A simple example would be the ISBN number for a book stored in an archive in electronic form.
- *Fixity* — Information needed to validate or authenticate the content information. This could be something as simple as a checksum that verifies that the bits within a CT image have not been corrupted since the image was produced. For sensitive data, this could include information needed to encrypt and decrypt the content information.

When a SIP is ingested into an archive, it is stored as one or more *Archive Information Units*. An example of an AIU might be a CT scan. Frequently AIUs are aggregated within the archive into *Archive Information Collections*. An example would be all CT scans for a single patient. And AICs can be aggregated into larger AICs: for example, all CT scans of all patients seen during a given year. As part of the submission planning process, archivists determine how SIPs will map into AIUs and AICs. There can be one-to-one, many-to-one, one-to-many, and many-to-many mappings between SIPs and AICs. Here are some examples:

- *One SIP - One AIC* — A government agency is ready to archive its electronic records from the previous fiscal year. All of the year's records are placed onto magnetic tapes that are submitted as one SIP. The archive stores the tapes together as a single AIC, with each tape identified as an AIU.
- *Many SIPs - One AIC* — A satellite sensor makes observations of the Earth over a period of one year. Every week all of the latest sensor data are submitted to the archive as a SIP. The archive has a single AIC containing all of the sensor's observations and the latest SIP is merged into that AIC.
- *One SIP - Many AICs* — A company submits financial records to an archive as one SIP. The archive chooses to store this information as two AICs: one which contains public information and the other which contains sensitive information. This makes it easier for the archive to manage access to the information.

- *Many SIPs - Many AICs* — An oil and gas company collects information on its wells. Every year it submits a SIP containing all of the well status information to an archive. The archive maintains one AIC for each oil or gas field and breaks out the information on each well to the proper AIC based upon its geographic coordinates.

A SIP will typically consist of a set of files transferred to the archive either over a network or on storage media. If the submission is on storage media, the archive may simply store the submitted media or copy the information to different media before placing the information under the control of the *Storage Entity*. When a SIP arrives at the archive via a network transfer, the information is recorded onto appropriate media and maintained thereafter by the Storage Entity. Over time, the media on which information is archived will degrade and may become obsolete, and the information will need to be copied to ensure preservation. This process is called *data migration*. There are three levels of data migration:

- *Replication* — Straight copying of data with no alteration in physical format, logical format or information content. An example would be copying a tape to the same type of tape.
- *Repackaging* — Migration wherein the physical or logical packaging surrounding the content information may change but the information content is unaffected. An example would be copying a reel-to-reel tape to a newer type of cartridge tape.
- *Transmutation* — Migration where the information content changes in format but (ideally) where information content is preserved. An example would be converting character-coded data from ASCII to a newer format like Unicode (which includes ASCII as a subset).

The fundamental rule of data migration is that information content should be preserved if at all possible. For example, an image with 16-bit pixels probably cannot be changed to have only eight bits per pixel, because information would then be lost. On the other hand, the image could be compressed using a lossless compression algorithm with no diminishment in the information content. The archive's operating policies, along with the Submission Agreement, define what constitutes "preservation of information" for a particular AIC.

Typically the archive will retain all of the preservation information that was supplied in a SIP. This information will usually be attached to the pertinent AIUs and AICs and maintained by the Storage Entity, but some of this information may be maintained elsewhere in the archive to facilitate the management and retrieval of information. For example, if a company archives the CAD drawings for its products, it may keep a database that identifies each product by type, model name and number, dates of manufacture, and so on. This *catalog information* can make it easier for engineers to locate drawings of interest within the archive. It is up to the archivists to determine the kinds of catalog information to maintain; however, the community that an archive serves may have certain standards or guidelines regarding catalog format, content, or access methods that should be followed to allow users to locate information that may be distributed across several sites.

Up to this point our scenario has discussed the submission of information as SIPs and the

maintenance of information within the archive as AIUs and AICs. We turn our attention now to the retrieval of archived information. The chief reason that SIPs, AIUs, and AICs all include preservation information is to assure that the content information can be retrieved at a later (possibly much later) date. Typically the retrieval of information from an archive is a two-step process:

- First, the information of interest must be located and ordered. In the OAIS functional model, this is supported by the *Access Entity*.
- Second, the information must be extracted from storage, packaged, and delivered. These functions are performed by the *Dissemination Entity*.

As an example of this process, consider a physician who is doing research on a certain type of tumor. She has a list of patients who were diagnosed with this malady and she wants to see if the hospital archive has their CT scans, and if so she wishes to retrieve them. She connects to the hospital archive through the Access Entity which begins by validating that she is authorized to access this information. She then enters a query to see if there is information in the archive for any of the patients on her list. Different archives will handle queries like this in different ways. The Access Entity may have to retrieve from storage an AIC that holds all CT scans and search through all of the collection's AIUs looking for those that qualify. Alternatively the archive may have a catalog that provides a summary of each scan, including the patient's name. Having such a catalog would greatly expedite the search in our example, since scans of interest could be located without having to sift through the archived information.

Once the full set of qualifying CT scans has been identified, the Dissemination Entity takes over to transfer the results to the consumer. There are several options for disseminating the information to the consumer: the scans could be recorded onto a digital tape and sent through the mail; or they could be transferred over an electronic network.. Regardless of the method of distribution, a *Dissemination Information Package* is assembled. A DIP comprises the content information to be distributed — in this case the CT scans — along with some or all of the supplemental information that the archive contains on the scans. Just as archivists determine the mapping from SIPs to AICs, so too the archivists determine how to map from AICs to DIPs. First, the information in a DIP may be in the same format in which it is stored within the archive or it may undergo some reformatting —repackaging or transmutation, as discussed above — before it is sent to the customer. Second, a DIP may include the information from a single AIU or a collection of AIUs extracted from one or many AICs. In our example, only a small fraction of the total number of CT scans in the archive actually fits the consumer's query, so the archive prepares a DIP containing only the desired subset. Alternatively, if the archive determined that the desired scans resided on a single computer tape (along with many other non-qualifying scans), the archivists could simply choose to replicate the particular tape and send it the researcher along with package information that would specify how to locate the scans of interest on the tape.

*** [End of document] ***